UNDERSTANDING THE FUNCTIONAL CONSEQUENCES OF GENETIC VARIATION ON GENE REGULATION

BY

FARZANEH KHAJOUEI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Professor Saurabh Sinha, Chair
Professor David Arnosti, Michigan State University
Professor Roy Campbell
Associate Professor Roy Dar
Associate Professor Jian Peng

ABSTRACT

Understanding the relationship between the non-coding sequence of the genome and the gene expression is one of fundamental goals of regulatory genomics. Perturbing certain locations in the non-coding DNA causes a disturbance to the precise spatial and temporal expression of the genes. Gene regulatory mechanisms determine the amount of change in the gene expression from variations in the sequence. Mathematical modeling of gene expression has been proven to be successful to establish a sequence-to-function relationship in a context aware manner and provide mechanistic explanation of the gene regulatory processes. In this thesis, we aspire to provide tools for understanding sequence-level encoding of gene regulation by applying thermodynamics-based models. More specifically, we provide a probabilistic framework to develop deeper insights about current knowledge of a gene's regulatory mechanisms, objectively characterizing what a new experiment adds to such knowledge and quantifying how 'informative' that experiment is. In order to elucidate mechanisms of transcriptional regulation we use single nucleotide polymorphism data to further investigate different mechanistic hypotheses and provide knowledge of systems-level processes. We construct a probabilistic model to leverage our knowledge of transcriptional regulatory networks and identify variations that lead to a significant change. Through this work we not only advance the field of regulatory genomics, but potentially provide a path for identifying variations in the DNA that significantly effect phenotype and lead to a disease.

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude towards my amazing mentor Professor Saurabh Sinha for his support and guidance through all steps of my academic progress. He helped me build confidence and patiently taught me all necessary skills for research. I would also like to thank the past and present members of Sinha lab research group: Charles Blatti, Thyago Duque, Hassan Samee, Shayan Tabe Bordbar, Payam Dibaei, Saba Ghaffarian, Xiaoman Xie and Shounak Bhagole. I would like to thank Casey Hanson, Laura Sloofman and Pei-Chen Peng for all great discussions and moral supports. Additionally, I would like to thank Bryan Lunt for interesting scientific discussions.

Secondly, I would like to thank my friends and family who supported me through all ups and downs of the journey. Specially my partner and companion Fred Douglas who always encouraged me and made my everyday life brighter. I am grateful for having wonderful friends Maryam Khademian, Parisa Hosseinzadeh, Behnaz Arzani, Shohreh Shaghaghian, Samaneh Mesbahi and Nasrin Sarrafi who always believed in me and provided a strong emotional support network. Lastly, I want to thank my cat, Oolong, for being a great companion and co-author of this thesis.

To my parents, for their love and support.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 Concepts in Gene Regulation 1
1.2 Thermodynamics Based Sequence-to-Expression Models 2
1.3 Ensemble Approach to Understand the Mechanisms of Gene Regulation . . . 4
1.4 Quantifying the Effect of Single Nucleotide Variations to Distinguish Mechanistic Models of Gene Expression 5

CHAPTER 2 AN INFORMATION THEORETIC TREATMENT OF SEQUENCE-TO-EXPRESSION MODELING 7
2.1 Ensemble of Models to Predict Gene Expression 7
2.2 Construction of a Probability Distribution over Models, and Potential Applications 10
2.3 An Information Theoretic Measure of the 'Value' of an Experiment 14
2.4 Evaluating Perturbation Experiments on *ind* Gene Regulation, Ex Post Facto 17
2.5 Interpreting the Information Gained from an Experiment 21
2.6 Quantifying and Interpreting the Value of Perturbation Experiments on the *sim* Enhancer 22
2.7 Detailed Description of the Ensemble Analysis 23
2.8 Discussion 28

CHAPTER 3 MODEL-BASED ANALYSIS OF POLYMORPHISMS IN AN ENHANCER REVEALS CIS-REGULATORY MECHANISMS 31
3.1 Understanding the Gene Regulatory Mechanisms Using Ensemble of Models 31
3.2 Thermodynamics-Based Modeling of Gene Regulatory Mechanisms 32
3.3 Statistical Testing of Compensatory Effects of SNPs 33
3.4 Expression Data Support Diverse Mechanistic Models of *ind* Enhancer Function 34
3.5 Model-based Analysis of Polymorphisms in the *ind* Enhancer 38
3.6 Experimental Testing of Selected Polymorphisms Identifies a Single Mechanism 42
3.7 Final Ensemble Predicts the Gene Expression in Orthologous Enhancers and Variants of *rhomboid* Enhancer 44
3.8 Discussion 48

CHAPTER 4 TISSUE SPECIFIC ANALYSIS OF NON-CODING VARIATIONS 51
4.1 Studying Genetic Variations in the Non-Coding DNA 51
4.2 Analysis of SNPs Leveraging Gene Regulatory Evidence 53
4.3 A Probabilistic Model to Integrate Regulatory Evidence with Expression Quantitative Trait Locus 56
4.4 Using the Probabilistic Graphical Modeling to Prioritize SNPs in Simulated and Real Data 69

4.5 Discussions 73
CHAPTER 5 REFERENCES 74
APPENDIX A SUPPLEMENTARY FIGURES AND TABLES 84

CHAPTER 1: INTRODUCTION

Rapid advances in sequencing technologies promise to lead us to a new era of personalized diagnosis and treatment based on genomics [1]. One of the most pressing challenges in this era will be to uncover the cellular and physiological impacts of DNA variations [2, 3], especially non-coding variations. Understanding gene regulatory mechanisms is critical in interpreting the genetic variations and their effects on human health [4]. The non-coding region harbors the majority of DNA variations that are associated with complex traits and common diseases, such as cancer, diabetes and Alzheimer's disease [5]. The most challenging task for researchers is to uncover the perturbed biological functions implicated by these DNA variations. A vast majority of variations associated with complex traits and common diseases fall in non-coding regions of the genome [5, 6, 7], and they can potentially impact gene regulatory functions [8, 9, 10]. While statistical genetics approaches have proven invaluable in short-listing such variants in specific disease contexts [11], to elucidate the mechanistic basis of these variants one needs a detailed quantitative understanding of regulatory sequence function that is often beyond state-of-the-art [12]. The field of computational genomics aims to understand the non-coding region of the human genome by developing statistical and mathematical models. The non-coding DNA contains sequences that act as regulatory elements influencing the transcription of genes into protein products through a process called transcriptional regulation.

1.1 CONCEPTS IN GENE REGULATION

The Human Genome Project (HGP) made history in 2004 by publishing the first high quality map of a human genome (the germline DNA sequence passed from parents to offspring via reproduction). This historic achievement came at the significant cost of $2.7 billion over 13 years and $100 million for a single genome in 2001. Since then, the cost of whole genome sequencing (WGS) per genome has decreased dramatically year over year, to the point where, in 2019, it costs just $1000 - a staggering $10,000$ fold reduction over the 2001 price [13]. Aside from the biological utility for WGS increasing demand and driving cost down, the main appeal underlying the demand for this technology lies in its potential to revolutionize medicine; physicians could be empowered to make informed and precise medical decisions on diagnoses and treatment on the basis of the patient's unique genetic makeup, thereby avoiding wasteful periods of trial and error that incur high cost, congest the health system, and can prove tragically fatal. Furthermore, the low cost of the technology and the ubiquity

of its adoption could facilitate the discovery of even more consequences of particular genetic variants, leading to even better standard of care. Of course, to ascertain the consequence of an individual's given genetic makeup requires data variation in genetic makeup; thankfully, the observed diversity emerges from variation in genomes at the genetic level.

The composition of the DNA sequences are two intertwining chains of bases (nucleotide) of which there are only four kinds (Adenine - A, Guanine - G, Cytosine - C, Thymine - T); the chain is mostly identical across individuals but varies at particular positions in the chain - variations which we call genetic. The range of changes for genetic variations could be wide: from a single base pair change in the DNA that we refer to as single nucleotide polymorphism (SNP), to deletions or insertions of multiple nucleotide (indels) or larger structural variations such as copy number variation and chromosomal rearrangement events.

Cellular processes are determined by the response of regulatory sequences in DNA to signals from specific proteins called transcription factors (TFs), leading to up- or down-regulation of gene expression [14]. A major class of regulatory sequences is that of cis-regulatory modules (CRMs, also called enhancers): these are regions of DNA, about 500-2000 base pairs long, harboring TF binding sites that control the transcriptional levels of nearby genes. Variation of the DNA sequence in CRMs can affect gene expression, and has been linked to developmental defects and disease [5]. Even minor variations, such as single nucleotide polymorphisms (SNPs), in CRMs can have significant functional impact, such as problems in fetal development [15]. The consequences of sequence variations depend on various circumstances, such as the tissue that the gene is functional in, the regulatory network or functional pathways that the gene is involved in, developmental stage, and other factors [16, 7, 17]. Quantitative modeling of gene regulatory systems formalizes what is known about a gene's regulatory mechanisms and can predict the effect of variants in molecular levels [18, 19]. Mathematical models that are formulated and validated on biological data are capable of predicting the changed caused by variations to the sequence [20, 21].

1.2 THERMODYNAMICS BASED SEQUENCE-TO-EXPRESSION MODELS

Our ability to predict the impact of non-coding sequence variations on gene expression is very limited, in part due to the complexity of CRMs, and in part because such impact depends not only on the sequence itself but also the abundance and activities of relevant TFs in the cellular conditions of interest. Statistical methods based on correlations among diverse data types such as TF-ChIP, histone modifications, gene expression, etc. can reveal salient properties of CRMs such as their tissue-specific activities [22] and their major regulators [TFs] [23, 9], and can in some cases predict the effect of removing a TF's influence on a

CRM or gene [24, 25, 26]. Statistical and machine learning methods have recently been developed that can to some extent predict the effects of single nucleotide mutations on TF binding levels, DNA accessibility [27, 28], and even gene expression [29], but these are typically not amenable to mechanistic interpretations, and are in a relatively early stage of exploration.

On the other hand, biophysical models based on equilibrium thermodynamics that explicitly incorporate key interactions among TFs, DNA and the transcriptional machinery have proven powerful for mechanistic understanding of the gene regulation process [18]. Thermodynamics-based modeling of gene expression reveals the precise mapping between CRM sequence and the associated gene expression in a variety of cellular contexts, the so called 'readout' of the CRM. These models provide a means to formalize our assumptions about a CRM's cis-regulatory logic, especially how its functional elements combine to regulate a transcriptional output [30, 31, 18, 32, 33, 34]. They can generate predictions that can be empirically tested [35], e.g., by targeted misexpression or mutagenesis experiments. Indeed, they have been used to predict effects of site mutations [20], and also promise to provide precise, mechanistically grounded predictions of the effect of minor sequence changes in CRMs [35]. Furthermore, these models can reveal ambiguities in our mechanistic knowledge about a system given existing data; pinpointing these ambiguities helps with choosing the future experiments that would best improve knowledge of the system. The success of thermodynamic models has been demonstrated in the context of systems with high-resolution gene expression measurements, such as early-stage Drosophila embryonic development [31, 18, 32, 36].

1.2.1 Scaling the Gene Expression Level Measurement for TFs and Target Genes

Scaling of gene expression: The thermodynamics-based calculation of the GEMSTAT model is given by the formula $\frac{Z_{on}}{Z_{on}+Z_{off}}$ (Equation 1 in He et al. [36]). This formula represents the fractional occupancy of the basal transcriptional machinery (BTM), but not the gene expression level. The expression of a gene in a cell type is merely proportional to this fractional occupancy for that cell type. (Note: This is a feature of the Shea-Ackers model, from which the GEMSTAT model was derived. The transcriptional initiation rate is assumed proportional to the BTM occupancy, which then allows us to show that the equilibrium mRNA level must also be proportional to that occupancy.) In other words, GEMSTAT is incapable of modeling absolute expression levels. The modeling framework deals with this by either defining the accuracy of the model (goodness of fit) as the correlation between predicted and true expression levels across different cell types, or by assuming an unknown

constant of proportionality that must be multiplied with the predicted expression before computing its deviation from true expression. (In the present work, we used the latter approach.) In either case, the scaling of the gene expression levels before presenting them to the model is irrelevant, and our results will be unchanged if we did not scale the expression values to the range 0-1.

1.2.2 Scaling of TF Expression

GEMSTAT was designed to handle the variation of relative TF concentration levels across cell types (e.g., positions along an embryonic axis), since this is what is commonly available. The biophysics of TF-DNA interactions, on the other hand, needs to be formulated in terms of the absolute TF concentration in a cell. Thus, where the model uses the term $[TF]$ (concentration of TF in cell), we replace it with the term $\nu[TF]_{rel}$, where is the relative TF concentration in that cell. This leads to a free parameter ν for each TF, but it also means that the $[TF]_{rel}$ values can be input on an arbitrary scale. Thus our "scaling" the TF expression level to a range of 0-1 is not necessary, and is merely a convenience for visualization.

Furthermore, it turns out that this additional free parameter per TF does not ultimately add to the model complexity, for the following reason. The key formula modeling TF-DNA interactions in GEMSTAT is the formula for the "statistical weight" of a site S for a TF, whose optimal site is S_{max}, given by: $q(S) = K(S_{max}\nu[TF]_{rel}e^{LLR(S)-LLR(S_{max})})$ (Equation 2 in He et al. [36]). Here, $K(S_{max})$ is the unknown TF-DNA binding constant for the optimal site , S_{max}, $\nu[TF]_{rel}$ represents the TF concentration, and $LLR(S)$ is the traditional LLR score of site S. Note that the TF-DNA binding constant $K(S_{max})$ must be made a free parameter for each TF. Note also that the two unknown constants $K(S_{max})$ and ν only occur as a product and never separately, so for model inference purposes their product $K(S_{max}) \times \nu$ can be treated as a single free parameter. This product is in fact the free parameter that we refer to as the 'DNA binding parameter' in the paper. Thus, the parameter ν that scales the relative TF concentration to an absolute level does not cause an additional penalty in terms of model complexity.

1.3 ENSEMBLE APPROACH TO UNDERSTAND THE MECHANISMS OF GENE REGULATION

In-depth studies of gene regulatory mechanisms employ a variety of experimental approaches such as identifying a gene's enhancer(s) and testing its variants through reporter assays, followed by transcription factor mis-expression or knockouts, site mutagenesis, etc.

The biologist is often faced with the challenging problem of selecting the ideal next experiment to perform so that its results provide novel mechanistic insights, and has to rely on their intuition about what is currently known on the topic and which experiments may add to that knowledge. We seek to make this intuition-based process more systematic, by borrowing ideas from the mature statistical field of experiment design. Towards this goal, we use the language of mathematical models to formally describe what is known about a gene's regulatory mechanisms, and how an experiment's results enhance that knowledge. We use information theoretic ideas to assign a 'value' to an experiment as well as explain objectively what is learned from that experiment. We demonstrate use of this novel approach on two extensively studied developmental genes in fruitfly. We expect our work to lead to systematic strategies for selecting the most informative experiments in a study of gene regulation.

1.4 QUANTIFYING THE EFFECT OF SINGLE NUCLEOTIDE VARIATIONS TO DISTINGUISH MECHANISTIC MODELS OF GENE EXPRESSION

The non-coding genome exhibits a hierarchical organization of structural and functional units, including large topologically associating domains or TADs at the megabasepair-scale [37], accessible regions and enhancers at the kilobasepair-scale [38, 39] and transcription factor (TF) binding sites at the basepair-scale. It is believed that a common mechanism of variant impact on cellular function is by affecting TF binding site strength, and consequently the gene expression level driven by an enhancer [40, 41]. Thus, to investigate such mechanisms we need a precise quantitative method to predict expression level from enhancer sequence, i.e., a "sequence-to-expression" model [30, 31, 32, 18, 33, 34, 36]; such methods must be sensitive enough to predict the regulatory effect of relatively minor changes in enhancer sequence, as is often the case with individual variations.

It is challenging to predict the impact of small genetic changes such as single nucleotide polymorphisms on gene expression, since mechanisms involved in gene regulation and their cis-regulatory encoding are not well-understood. Recent studies have attempted to predict the functional impact of non-coding variants based on available knowledge of cis-regulatory encoding, e.g., transcription factor (TF) motifs. In this work, we explore the relationship between regulatory variants and cis-regulatory encoding from the opposite angle, using the former to inform the latter. We employ sequence-to-expression modeling to resolve ambiguities regarding gene regulatory mechanisms using information about effects of single nucleotide variations in an enhancer. We demonstrate our methodology using a well-studied enhancer of the developmental gene intermediate neuroblasts defective (ind) in *D. melanogaster* . We first trained the thermodynamics-based model GEMSTAT to relate the neuroectodermal

expression pattern of *rho* to its enhancer's sequence, and constructed an ensemble of models that represent different parameter settings consistent with available data for this gene. We then predicted the effects of every possible single nucleotide variation within this enhancer, and compared these to SNP data recorded in the Drosophila Genome Reference Panel. We chose specific SNPs for which different models in the ensemble made conflicting predictions, and tested their effect in vivo. These experiments narrowed in on one mechanistic model as capable of explaining the observed effects. We further confirmed the generalizability of this model to orthologous enhancers and other related developmental enhancers. In conclusion, mechanistic models of cis-regulatory function not only help make specific predictions of variant impact, they may also be learned more accurately using data on variants.

CHAPTER 2: AN INFORMATION THEORETIC TREATMENT OF SEQUENCE-TO-EXPRESSION MODELING

Studying a gene's regulatory mechanisms is a tedious process that involves identification of candidate regulators by transcription factor (TF) knockout or over-expression experiments, delineation of enhancers by reporter assays, and demonstration of direct TF influence by site mutagenesis, among other approaches. Such experiments are often chosen based on the biologist's intuition, from several testable hypotheses. We pursue the goal of making this process systematic by using ideas from information theory to reason about experiments in gene regulation, in the hope of ultimately enabling rigorous experiment design strategies. For this, we make use of a state-of-the-art mathematical model of gene expression, which provides a way to formalize our current knowledge of cis- as well as trans- regulatory mechanisms of a gene. Ambiguities in such knowledge can be expressed as uncertainties in the model, which we capture formally by building an ensemble of plausible models that fit the existing data and defining a probability distribution over the ensemble. We then characterize the impact of a new experiment on our understanding of the gene's regulation based on how the ensemble of plausible models and its probability distribution changes when challenged with results from that experiment. This allows us to assess the 'value' of the experiment retroactively as the reduction in entropy of the distribution (information gain) resulting from the experiment's results. We fully formalize this novel approach to reasoning about gene regulation experiments and use it to evaluate a variety of perturbation experiments on two developmental genes of D. melanogaster. We also provide objective and 'biologist-friendly' descriptions of the information gained from each such experiment. The rigorously defined information theoretic approaches presented here can be used in the future to formulate systematic strategies for experiment design pertaining to studies of gene regulatory mechanisms[1].

2.1 ENSEMBLE OF MODELS TO PREDICT GENE EXPRESSION

Mechanisms influencing the precise function of a regulatory system include the number, accessibility, affinities and relative arrangement of TF binding sites within a CRM, as well as cellular concentrations of the TF molecules, and protein-protein interactions; all of these mechanisms affect the rate of transcription of the gene [43]. Thermodynamic models of CRM function encode these mechanistic factors in their parameters, which correspond to biochemical properties of the molecules controlling the gene expression. These parameters are typically computationally optimized to be assigned values that can best explain the gene

[1]This work has been published in PLoS Computation Biology [42].

expression patterns attributable to a set of CRMs [18, 36]. When used to investigate the regulatory function of a single CRM, the thermodynamic modeling approach faces a significant challenge: non-uniqueness of the optimal models. For instance, a CRM mediating control by five or more TFs will include 10 or more free parameters in the GEMSTAT thermodynamic model [36] and we have shown previously that parameter training will converge to one of many local optima [20]. Each optimal model explains the data equally well, but uses parameters that correspond to significantly varying, often mutually incompatible mechanistic hypotheses [44, 45]. For example, consider a gene regulated by a CRM that is under the control of two activators. Assume there are two models that explain the wild-type expression pattern. One model predicts the correct expression by using (assigning function to) only one activator, while the other model uses both activators. In the absence of additional biological experiments that confirm the role of each activator, both models are equally plausible. Problems arise when we try to predict, using the model, the effect of knocking down an activator or mutating its binding site(s). Depending on which model we use, the predicted effect of the perturbation experiment is different: a model that does not use an activator will not predict a change due to removal of that activator's influence. However, there is no reason to prefer one model's prediction over the other, and the biology remains ambiguous until a new experiment is performed. We believe it is important to respect this ambiguity of knowledge when modeling gene expression data and making predictions about future experiments.

In agreement with the above proposal, Samee et al. [20] laid out a new paradigm of gene expression modeling where one searches within the model's parameter space for as many optima as possible, resulting in an "ensemble" of optimal models. (Henceforth, different assignment of values to the model's tunable parameters will be considered as different models.) Each model in the ensemble is a hypothesis about the cis-regulatory mechanisms encoded in the CRM, and is also capable of making specific predictions about perturbation experiments. A simple approach to working with an ensemble of models is to make predictions by uniformly aggregating predictions of its member models. It has been shown that this "wisdom of crowds" approach can be effective: aggregated votes of many models can predict the effect of site mutations more accurately than any individual model [20].

We noticed, however, that a typical ensemble of sequence-to-expression models, e.g., that created by Samee et al. [20] in modeling embryonic expression of the Drosophila gene ind, is not uniformly distributed in the parameter space. Rather, they are clustered in the parameter space (Figure 2.1D), with models within a cluster predicting similar effects for a particular perturbation, but each cluster's consensus predictions being qualitatively different from those of other clusters. Different clusters can have different 'spans', i.e., the extent to which models in that cluster differ from each other quantitatively (in parameter values)

while producing equally good fits to available data and essentially the same predictions for future experiments. For instance, the cluster at the bottom of Figure 2.1D has greater span than the cluster on the top-left, which is relatively tight. The span of a cluster pertains to parameter sensitivity [44, 45] in that region of parameter space. Furthermore, different clusters may have different representation (number of models) in the ensemble, and the number of represented models may not correlate with the span of the cluster. This is because we do not make strong assumptions about how the ensemble of models was obtained, beyond that it is a collection of models that fit the available data and may be located in different regions of parameter space. With these observations about ensembles of models, we sought the most appropriate way to use ensembles for making predictions and for designing future experiments. We describe here one such procedure that we developed and implemented, which allows us to make predictions with ensembles of models, and also offers a principled approach to experiment design in gene regulation studies.

Briefly speaking, our modeling approach involves (1) creating a large ensemble of models that fit the available data accurately, following the sampling and optimization strategy of Samee et al [20] and (2) defining a probability distribution over the parameter space such that the ensemble of models represents regions of high probability and where each cluster of models (roughly speaking, a distinct mechanistic hypothesis) has approximately the same total probability as other clusters. This distribution provides a principled way for us to make aggregated predictions about any particular perturbation experiment, and to describe the uncertainty in such predictions. Additionally, we show how to measure the entropy of this probability distribution, thereby quantifying the uncertainty in parameter space [46] that remains after fitting the models to available data. Noting that the ensemble of models consistent with available data changes (typically shrinks) upon performing an additional experiment, we suggest that the difference of entropies of the probability distributions before and after an experiment (i.e., information gain) may be used to score the 'value' of the experiment. We can use this value as a score to compare different experiments, the experiment with greater score being deemed the more informative experiment. The ability to assign information theoretically-grounded 'values' to experimental results is significant, since it paves the way for principled experiment design [47, 48]

2.2 CONSTRUCTION OF A PROBABILITY DISTRIBUTION OVER MODELS, AND POTENTIAL APPLICATIONS

2.2.1 Outline of Gene Expression Model

We consider the class of mathematical models that predict the gene expression level driven by a cis-regulatory module (CRM) from the latter's sequence, given prior knowledge of relevant transcription factors (TFs), their in vitro DNA-binding affinities (motifs), and their concentration levels in the cellular context of interest. Several such models have been investigated in the literature [31, 18, 32, 35], and we work with the GEMSTAT model [36], which we developed previously and which we are most familiar with. The GEMSTAT model has two free (tunable) parameters for each relevant TF, one corresponding to its binding strength for the consensus site and one corresponding to its potency as an activator or repressor. The model also has optional free parameters for any TF-TF cooperative interactions that the modeler may choose to include. Assigning values to these free parameters specifies a model completely, allowing it to predict gene expression in any cellular context where TF concentrations are known. Typically, optimization strategies are used to identify the parameter setting(s) that accurately predict gene expression driven by a CRM in multiple cellular contexts [49].

2.2.2 Construction of Model Ensemble

In light of the observations made in Introduction, we sought to first construct an ensemble of models that are widely spread in parameter space, and thus represent different mechanistic explanations of data. A model is included if its goodness-of-fit score – sum of squared errors or 'SSE' between known and predicted expression levels in multiple cellular conditions – is below a threshold. We noted that the number of TFs in common modeling scenarios is less than 10 [31, 18, 32, 35, 36], and the number of free parameters in the range of 10-20. This led us to consider uniform sampling of the parameter space as the first step of ensemble construction. We followed the approach of Samee et al. [20] (Figure 2.1 A-D), performing extensive uniform sampling from the space (millions of samples), followed by filtering of promising models (SSE score below a modest threshold), local optimization seeded by these promising models, and a final round of filtering on the optimized models (SSE score below a strict threshold). (See Methods for details.) This procedure allows us to construct a large ensemble of models representing many or all optimal regions of the parameter space. We provide more details of ensemble size and composition later (specific gene models).

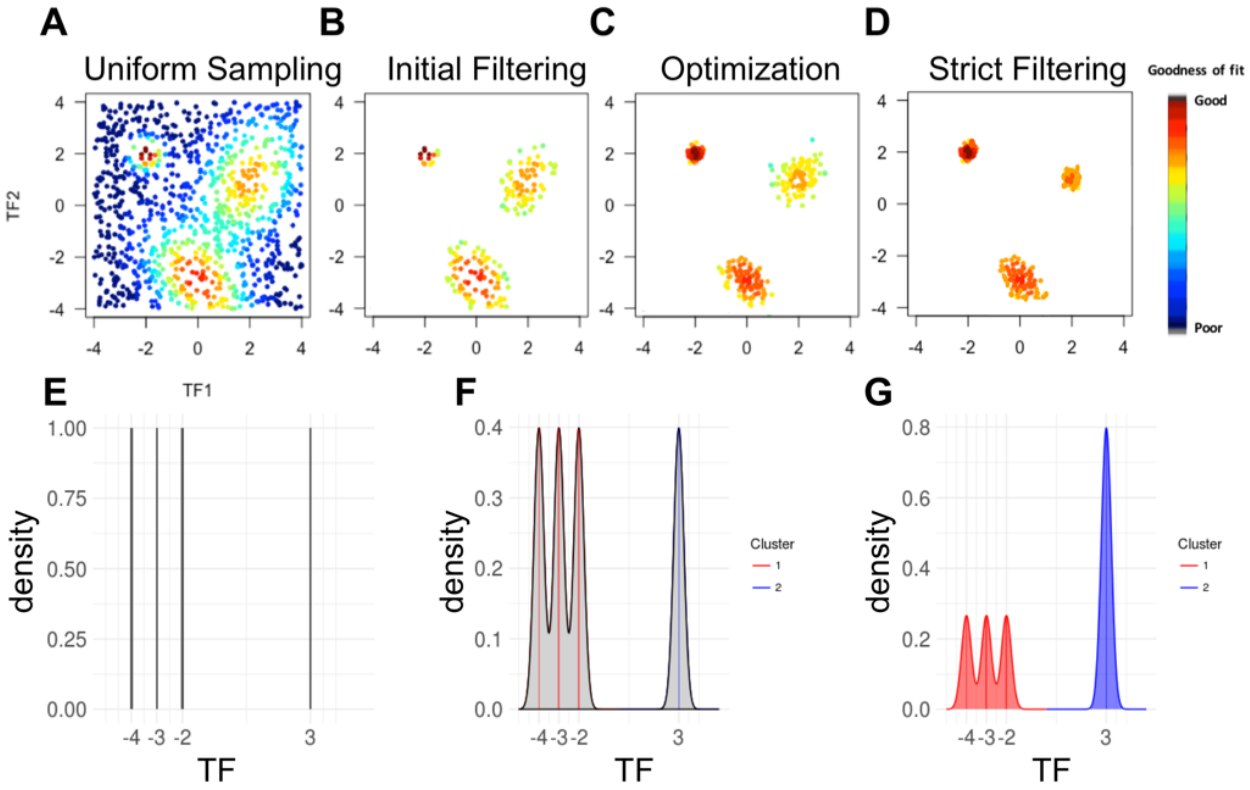

Figure 2.1: Schematic of Methodology (A) Each point in the scatter plot is a model with two parameters denoted TF1 and TF2 and its color shows its 'goodness-of-fit' score (see color legend on right). A number of models are uniformly sampled from the two-dimensional parameter space and scored against available data. (B) Initial filtering of models in (A) for a high goodness-of-fit score results in multiple clusters of models. (C) Models in panel B are used as starting points for numeric optimizations of goodness-of-fit, resulting in multiple clusters of locally optimal models. (D) A stringent threshold on goodness-of-fit is used to filter optimized models (panel C) to obtain the final ensemble. (E) An ensemble of four models in a one-dimensional space is represented by a uniform discrete probability distribution. (F) The same ensemble as (E), represented by a continuous probability distribution, as a uniform mixture of Gaussian distributions centered at the four models. (G) Models are recognized to fall in two clusters and the continuous distribution of (F) is modified to assign equal weight to each cluster.

2.2.3 Construction of Probability Distribution over Models

An ensemble of models can be used to make predictions by aggregating (averaging) the predictions made by each member model. However, this approach ignores the fact that the ensemble construction (outlined above or by a similar method) likely results in some regions of parameter space being over-represented in the ensemble. Models belonging to the same region, i.e., proximal to each other in the parameter space, are presumed to represent qualitatively similar mechanisms of CRM function. Thus, the ensemble's aggregate predictions may be biased towards one or a few mechanistic hypotheses. We therefore sought a more nuanced way to aggregate model predictions, by defining a probability distribution over parameter space that captures how the fit models are spread across different regions of the space but discounts for unequal representations of (number of models in) different regions. Such a probability distribution can then be used to make predictions about new experiments and also to score the uncertainty of mechanistic explanations offered by the ensemble. We also note that constructing this distribution has close ties to the kernel density estimation problem [50] but is different because the ensemble is not a collection of IID samples drawn from the desired population.

The simplest distribution to consider is a discrete uniform distribution over the models in the ensemble; e.g., Figure 2.1E shows such a distribution over an ensemble of four models in a toy 1-dimensional parameter space. In the continuous parameter space, highly proximal models are likely to have similar goodness-of-fit, therefore we smoothen the discrete distribution by centering a Gaussian distribution at each model in the ensemble and constructing a uniform mixture (Figure 2.1F). This mixture of Gaussian distributions provides a continuous distribution, but if one region of the space is over-sampled in the ensemble, the distribution puts undue weight in that region; e.g., the three closely-related models on the left in Figure 2.1F together carry about three times the probability mass as that around the isolated model on the right. In light of this observation, we first cluster models, each cluster roughly corresponding to a distinct mechanistic hypothesis, and define the overall probability distribution to be a mixture of distributions representing each cluster. Since we lack any additional knowledge to prefer one cluster over another, we assume uniform mixture weights for the clusters. The probability distribution representing each cluster, in turn, is a mixture of Gaussian distributions whose means are the models in that cluster. Thus, Figure 2.1G shows a mixture of two distributions (red and blue) representing the two clusters, with the red distribution in turn being a uniform mixture of Gaussians centered on the three models in that cluster. Figure 2.2A shows a similar construction, now for a 2D parameter space, beginning with the given filtered ensemble $\theta_1, \theta_2, \ldots, \theta_M$, identifying three

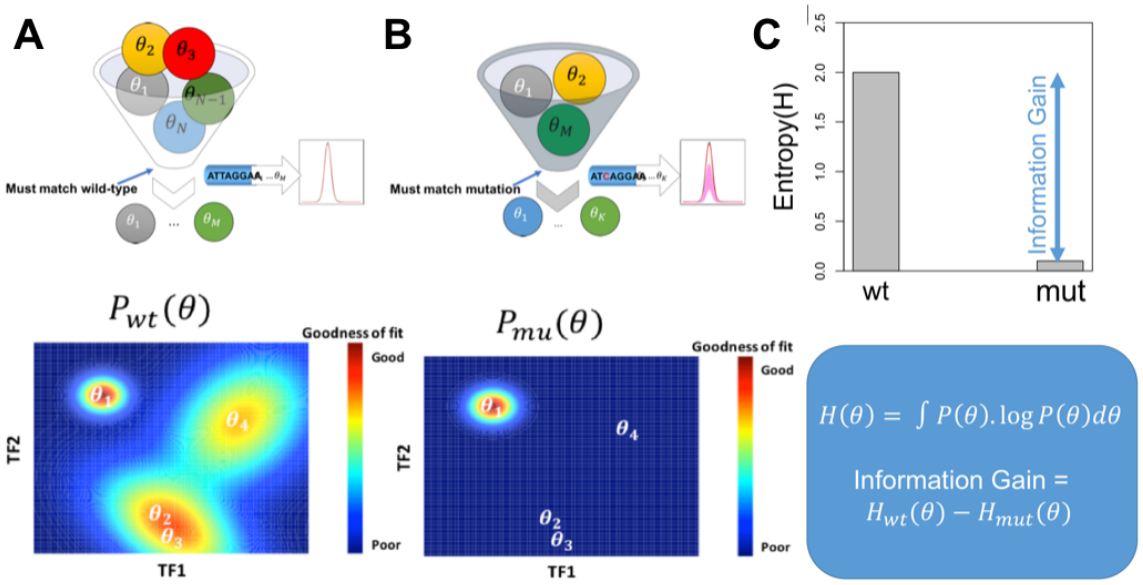

Figure 2.2: (A-B) Illustration of an information theoretic value of an experiment: (A) All models are assessed for goodness-of-fit on available (e.g., wild-type) data, an ensemble $\theta_1, \theta_2, \ldots, \theta_M$ is obtained (top) as outlined in panels A-D, and a probability distribution is constructed (bottom), shown here for a two-dimensional parameter space. (B) Models in the wild-type ensemble are further examined for goodness-of-fit on new experimental data (e.g., a site mutagenesis experiment), and fit models are retained to construct a new filtered ensemble (top) and the probability distribution is recomputed (bottom). (C) Entropy scores of the probability distributions of the wild-type ensemble and filtered ensemble are computed, denoted by H_{wt} and H_{mut} respectively, and the information gain is computed as the difference of these two entropy scores.

clusters and constructing the mixture probability distribution. (For more details, especially the construction of co-variance matrices for these distributions, see Methods.)

2.2.4 Potential Applications

The probability distribution over models, constructed as above, can be used in the following ways:

1. Aggregating ensemble predictions: Each model in the ensemble makes a prediction on unseen data, i.e., the gene expression level driven by the modeled enhancer(s) in a cellular context described by TF concentrations. The ensemble's aggregate prediction can be computed by averaging predictions of all models weighted by their probabilities as specified by the distribution.

2. Quantifying uncertainty of ensemble predictions: The probability distribution makes it possible to quantify the variance in ensemble predictions on unseen data. This variance represents the uncertainty among ensemble models with respect to that data point.

3. Assigning objective 'value' to an experiment: We can utilize the probability distribution over an ensemble to measure the information theoretic value of an experiment. Given new data, i.e., results from a new experiment, we can filter models to obtain a smaller ensemble that agrees with the new data, henceforth called the 'filtered ensemble' for the experiment. Using the fact that the entropy of the probability distribution captures the uncertainty intrinsic to the ensemble, we can use the difference in entropy of the original ensemble and the filtered ensemble, also called the 'information gain', as an objective evaluation of how informative the experiments results are. (Details of entropy calculation for a given probability distribution are provided in Methods, but see Figure 2.2A,B for an illustration.) This approach ultimately allows, though we do not explore it here, principled experiment design where the experiment whose results are expected to result in the greatest information gain, is selected for follow-up.

2.3 AN INFORMATION THEORETIC MEASURE OF THE 'VALUE' OF AN EXPERIMENT

Sequence-to-expression models enable us to propose mechanisms for gene expression regulation, that may then be confirmed by performing perturbation experiments such as TF knockout or site mutagenesis, followed by expression assays that inform us about how the gene expression changes in the perturbation condition. Some experiments result in greater gene expression changes than others, and it is natural to want to characterize 'what was learned' from each experiment, as well as quantify how informative that experiment was. Here, we demonstrate such an exercise in systematic experimentation in the gene regulation context, using the ensemble modeling framework described above.

Our first set of demonstrations are in the context of the regulatory mechanisms of an early development gene in D. melanogaster - the intermediate neuroblasts defective (ind) gene. We chose this gene because it is known to be regulated by a well-defined enhancer, and its major regulatory inputs are well characterized. The gene was characterized by Weiss et al. [51] and Stathopoulos and Levine [52], among others, and was the subject of systematic modeling by Samee et al. [20]. It is expressed in a lateral stripe along the dorso-ventral axis of the early embryo (S1A Fig, black curve), with activation from the TFs Dorsal (DL) and Zelda (ZLD), and repression by the TFs Snail (SNA), Ventral nervous system defective

(VND) and Capicua (CIC) (S1A Fig). In addition to the wild-type expression pattern of this gene, its expression has been experimentally recorded under several perturbation conditions (S1 Table), surveyed by Samee et. al [20] and further discussed below.

Despite the knowledge of a fairly complete set of regulatory inputs, several ambiguities remain about the cis-regulatory logic of the *ind* enhancer. This is evident when we construct an ensemble of models that predict the known expression pattern of *ind* from its enhancer sequence along with TF concentration profiles along the D/V axis. Figure A.1B shows that the ensemble's mean prediction (magenta curve) for these wild-type conditions fits the wild-type expression profile accurately, and with little variation among different models (pink curves are models in the ensemble) but Figure A.1 C reveals that most of the 13 parameters of the model exhibit substantial variability, a point also illustrated by the marginal distributions of ten of the parameters (S1D-F Figs). The high degree of uncertainty is not surprising, given that data from only one experiment – the wild type condition – for a single enhancer was used to train the ensemble. It also means that results of various perturbation experiments may prove informative about this gene's regulatory mechanisms, an avenue that we pursue next.

First, we worked with a 'synthetic true model' MST that allows us to predict results of various perturbation 'experiments' in silico. This synthetic true model MST was carefully chosen from among the ensemble of models consistent with wild-type data, described above. (See S3 and S4 Figs for details.) We used MST to individually predict the effects of (a) each TF's knockout and (b) removing the strongest site of each TF in the enhancer, and treated these predicted gene expression patterns (Figure 2.3 A, green curves) as the 'true' results of those hypothetical or 'in silico perturbation experiments'. We used each of the 10 in silico experiments to construct a 'filtered ensemble' (average predictions shown in Figure 2.3 A, magenta curves), computed its entropy score, and thus assigned an information theoretic 'value' to the experiment (Figure 2.3 B). We noted that the magnitude of change in the expression profile resulting from a perturbation experiment does not necessarily reflect the value of the experiment. For instance, it is possible to obtain new information from a perturbation experiment where the expression pattern remains unchanged from wild-type, a case in point being the SNA knockout experiment (Figure 2.3 A), with assigned value 1.66 - apparently many models consistent with wild-type data cannot explain this experiment and are removed in the filtered ensemble from its results. Conversely, an experiment with a more substantial expression profile change may not add anything to our knowledge of the regulatory mechanism. For instance, the DL knockout experiment shows peak *ind* expression diminishing by $\approx 60\%$ (Figure 2.3A) but is assigned a value of 0.32 (Figure 2.3B), among the lowest of the 10 experiments; this is because most models capable of explaining the wild-type

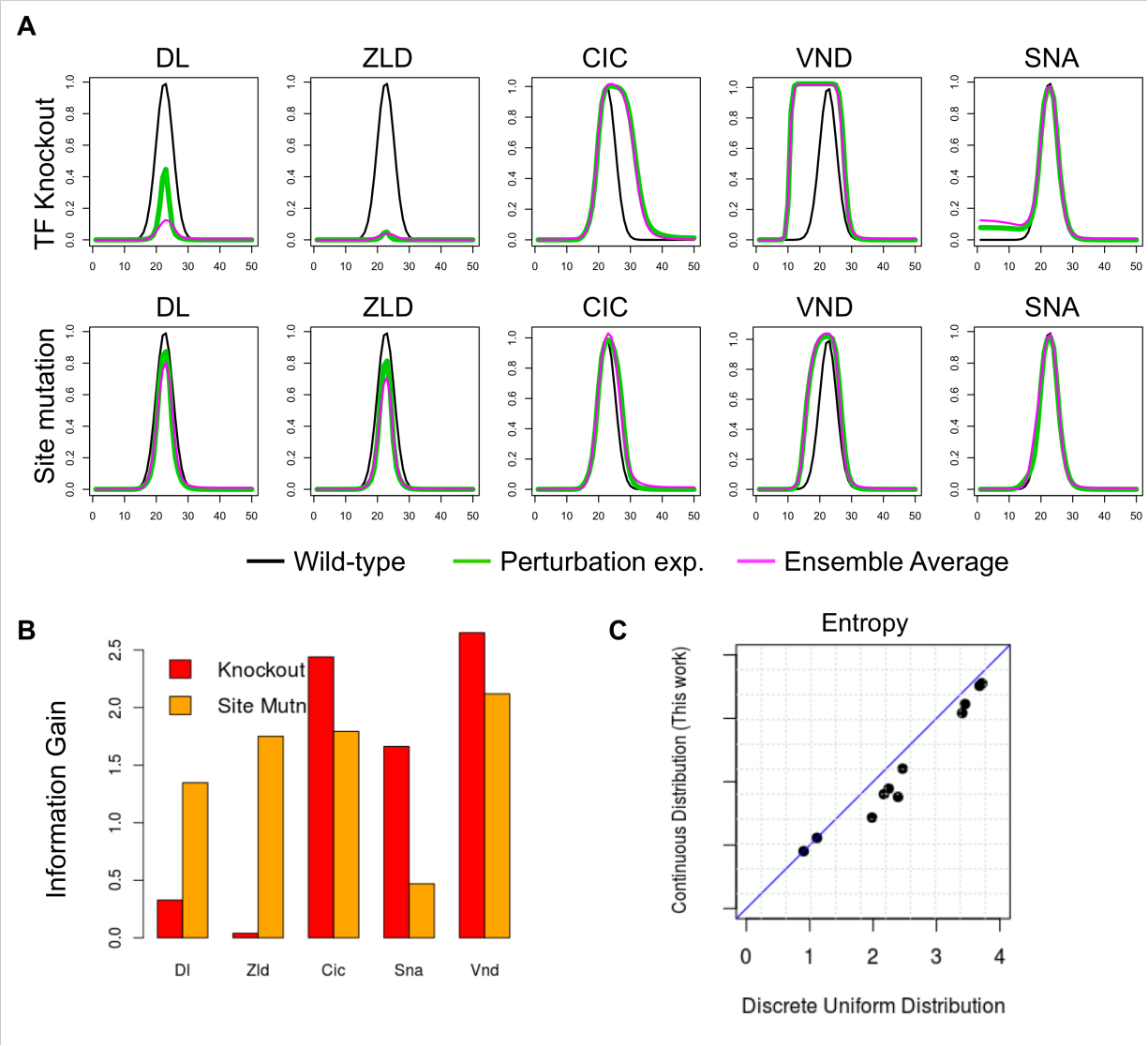

Figure 2.3: Evaluating in silico experiments with a 'synthetic real' model MST. (A) The model is used to generate synthetic 'experimental' results of TF knockout (top row) or strongest site mutagenesis (bottom row), for each TF, shown in green. These are compared to the synthetic 'wild-type' expression profile of ind, shown in black (in each panel). Magenta curves show the average prediction of the filtered ensemble for each of these 'experiments'. (B) Each of the ten synthetic 'experiments' (panel A) is assigned a value, which is defined as the information gain due to that experiment. (C) Entropy of filtered ensemble for each of the ten 'experiments', as defined under the specially constructed probability distribution presented in this work (Y axis) or under a discrete uniform distribution (X axis).

ind pattern apparently use DL as activator, so knocking out DL does not provide much new information. The same is not true of the experiment where the strongest DL site is removed, an experiment with minor impact on expression (Figure 2.3 A) but a relatively high assigned value of 1.34. This points out that even if the involvement of a TF is beyond doubt, there may be uncertainty regarding the strength of its regulatory input and the mediatory role of each of its binding sites. We noted (Figure 2.3 B) the same trend – that the value of strongest site mutagenesis is greater than that of TF knockout – for the other activator (ZLD). On the other hand, for perturbations involving repressors (SNA, VND, CIC) the value of the site mutagenesis experiment is less than that of TF knockout in all three cases. Also, for comparison, we show in Figure 2.3 C the relative values of the 10 'experiments' under a more simplistic scheme that evaluates each experiment by the reduction in entropy assuming a discrete uniform distribution on all models in an ensemble. We note that the two schemes largely agree with each other in this evaluation, though this may not be true in general, depending on how an ensemble of models is generated. Finally, we note that the observations above were made with a specific choice of the 'synthetic true model' MST, that furnished 'experimental' results, but the reported trends, e.g., large information gain from a perturbation experiment with little effect on expression, or little gain from an experiment with large effect, were unchanged when we repeated the entire exercise with a different choice of MST (S5 and S3B Figs).

2.4 EVALUATING PERTURBATION EXPERIMENTS ON *IND* GENE REGULATION, EX POST FACTO

In this section, we will examine results of real perturbation experiments pertaining to the *ind* gene reported in the literature and evaluate each experiment in the way described above. In addition to the wild type gene expression pattern of the *ind* gene (S1A Fig), we have information from six different biological perturbation experiments (S1 Table). It is known that *ind* expression is abolished in DL mutants [53] and becomes weaker in ZLD mutants [54]. Its peak expression reduces to 50% of its wild-type level upon mutation of the four strongest ZLD binding sites [20]. (We call this experiment 'ZLD site mut.'.) Removal of the strongest DL site ('DL 1 site mut.') has no observable effects on the expression [55] and removing three overlapping DL sites ('DL 3 site mut.') greatly diminishes peak expression [20]. Knockout of SNA (experiment 'SNA KO') leaves *ind* expression unaltered [32], while knocking out VND ('VND KO') causes the domain of expression to expand ventrally [56], and CIC site mutagenesis ('CIC site mut.') expands *ind* expression dorsally [57]. We evaluated each of the six perturbation experiments (two TF knockouts and four

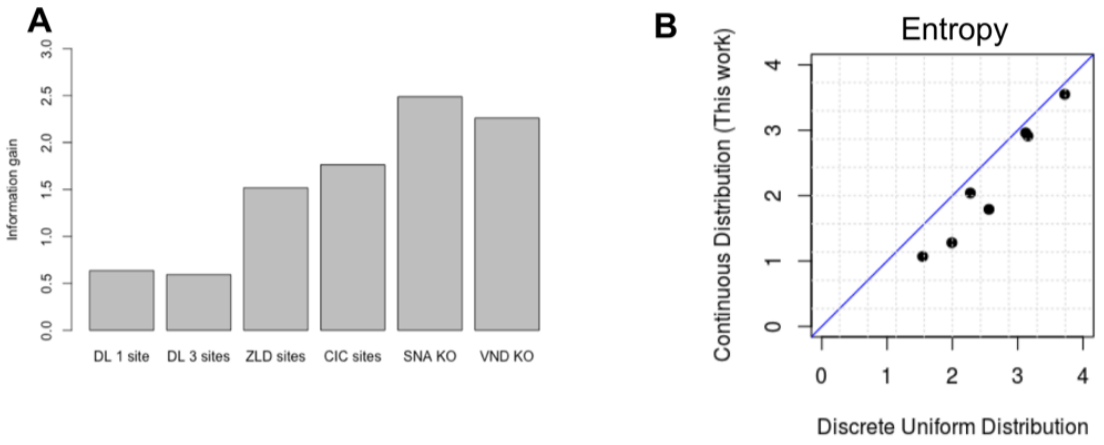

Figure 2.4: (A) Information gain score assigned to each of six perturbation experiments involving regulation of the *ind* gene. These include four site mutagenesis experiments (four left bars) and two TF knockout experiment (two right bars). (B) Entropy of filtered ensemble for each perturbation experiments, as defined under the specially constructed probability distribution presented in this work (Y axis) or under a discrete uniform distribution (X axis).

site mutagenesis experiments) using the approach introduced in the previous section – begin with the ensemble of models that explain wild-type gene expression, construct a filtered ensemble that additionally explains the perturbation results (see Methods and Figure 2.2), and calculate the difference in entropy ('information gain'). The values assigned to these experiments are shown in Figure 2.4A, and we note that the SNA and VND knockout experiments were the most informative in this group.

Evaluating a new experiment, in our scheme, involves ruling out from the original ensemble a subset of models inconsistent with that new experiment. Recall that models in the ensemble were clustered, with the informal understanding that each cluster represents a distinct mechanistic hypothesis. Thus, if an entire cluster is ruled out by a particular experiment, one may interpret it as ruling out a particular mechanistic hypothesis. Table 1 shows the sizes of clusters in the original (wild-type) ensemble of models and the effect of filtering with each perturbation experiment. We note that an experiment ('DL 3 site mut.' in Table 1) may remove just one cluster, while retaining other clusters of models as feasible. There may also be experiments ('SNA KO' and 'VND KO' in Table 1) that rule out the majority of mechanistic hypotheses, retaining only 2-3 of the original clusters. The other scenario – where all clusters are retained but rendered substantially sparser – is also seen, indicating that the information gained by those experiments was more along the lines of quantitative refinement rather than qualitative pruning of the space of possible mechanisms.

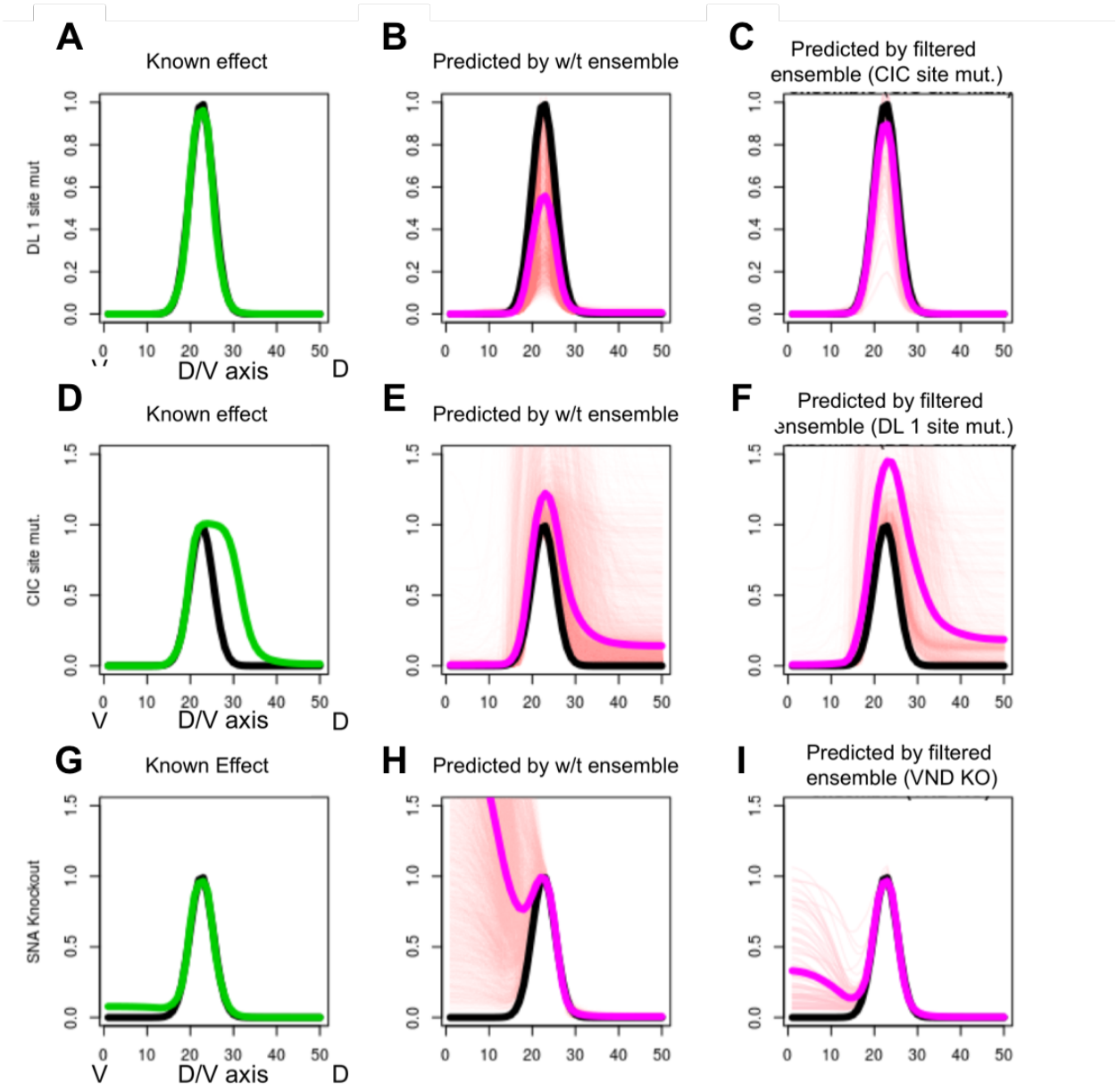

Figure 2.5: (A-C) Known effect of the 'DL 1 site mut.' experiment (A, shown in green, compared to wild-type in black) is better predicted by the filtered ensemble for the experiment 'CIC site mut.' (B) than by the wild-type ensemble (C). In B, C, ensemble predictions shown as thin red lines and their mean shown in thick pink. (D-F) Known effect of the 'CIC site mut.' experiment (D) is better predicted by the filtered ensemble for the 'DL 1 site mut.' experiment (E) than by the wild-type ensemble (F), which shows greater variance and disagreement across models. (G-I) The 'SNA KO' experiment has been observed to not affect *ind* gene expression pattern (G). In contrast, majority of the models in the wild-type ensemble (H) predict substantial ventral de-repression, but an ensemble filtered by the 'VND KO' experiment predicts far less change in the ventral domain of *ind* expression profile.

Figure 2.4B shows the above information theoretic evaluation of each experiment, compared to a simpler scoring scheme where entropy of an ensemble is simply the logarithm of the size of that ensemble, i.e., where we assume a uniform discrete distribution on models. As expected, the two scores are highly correlated.

Table 2.1: Models in the wild-type ensemble can be clustered into 8 different groups based on their parameter values, each cluster roughly corresponding to a distinct mechanistic hypothesis. Given information from a new experiment, we filter the wild-type ensemble for models that are predictive of the experiment outcome. Shown here are the sizes of clusters of models in the wild-type ensemble, and their corresponding sizes after filtering for each of six different perturbation experiments. The total number of models in each ensemble, the corresponding entropy score and information gain score are shown in the last three columns.

Ensemble	1	2	3	4	5	6	7	8	Total	Entropy	Information Gain
Wild-type	433	921	425	711	680	744	552	771	5237	3.55	0
DL 1 site mut.	111	239	112	169	203	215	286	98	1433	2.91	0.64
DL 3 site mut.	122	45	115	0	225	282	96	437	1322	2.96	0.59
ZLD site mut.	2	32	17	55	2	21	31	28	188	2.04	1.51
CIC site mut.	35	2	33	1	0	0	7	284	362	1.79	1.76
SNA KO	1	0	0	0	0	0	0	34	35	1.07	2.48
VND KO	3	0	1	0	0	0	0	94	98	1.29	2.27

Note that experiments were assigned values above under the assumption that they were the sole (or first) perturbation experiment performed. In reality, of course, a line of enquiry proceeds via a series of such experiments, begging the question whether a perturbation experiment can be informative on its own but not so much if it follows another perturbation experiment. We explored this question further, by examining every possible pair of experiments (performed sequentially), and noted that there are indeed such examples. However, in the interest of continuity we do not discuss this analysis here, referring the interested reader to Supplementary Table A.3.

2.5 INTERPRETING THE INFORMATION GAINED FROM AN EXPERIMENT

We next moved beyond asking 'how much' information was gained from an experiment to the more subjective question of 'what' information was gained. To answer this, it seems natural to compare the original (wild-type) ensemble of models to the filtered ensemble that is additionally consistent with the new experiment's results. The challenge then becomes: how do we compare these two ensembles in a language that appeals to the biologist's intuition? One pragmatic approach that we devised, and illustrate here, is to identify a second experiment for which the two ensembles make markedly different predictions, and use this difference to illustrate the distinction between ensembles. For instance, consider the 'CIC site mut.' experiment, which we saw above to be of modest information theoretic value (Figure 2.4A). We also noted in Table 1 that this experiment induces a filtered ensemble with two of the eight original clusters completely ruled out and two additional clusters drastically reduced in size (from 900 and 700 models to 2 and 1 models respectively), suggesting that certain plausible mechanistic hypotheses were indeed ruled out by it. To interpret this further, we considered the predictions of this filtered ensemble on the 'DL 1 site mut.' experiment (Figure 2.5C) and found these to be in fair agreement with the true results from the literature [55] (Figure 2.5 A). We then noted that the wild-type ensemble, not filtered by the 'CIC site mut.' experiment, is far more uncertain in its predictions about the 'DL 1 site mut.' experiment (Figure 2.5B). Thus, the 'CIC site mut.' experiment informs us, correctly, that mutagenizing the strongest DL site in the enhancer should not result in a significant reduction in peak *ind* levels, a point that was ambiguous in the original ensemble.

A similar approach can be adopted to interpret the information provided by other perturbation experiments. In our second example, we interpreted the 'DL 1 site mut.' experiment by examining the predictions of its filtered ensemble on the 'CIC site mut.' experiment, which according to the literature [57] shows an extension of the dorsal boundary of *ind* expression (Figure 2.5D) This derepression effect is much more accurately predicted by the filtered ensemble (Figure 2.5E), while the original ensemble's average prediction is less definitive in predicting this effect (Figure 2.5E). In other words, the 'DL 1 site mut.' experiment informs us that CIC is an important repressor of the *ind* gene, setting up its precise dorsal boundary. For our third example, we note that the filtered ensemble of the 'VND KO' experiment accurately predicts that a genetic knockout of SNA will not affect the ventral boundary of *ind* expression (Figure 2.5F,G), while the original ensemble erroneously predicts ventral de-repression (Figure 2.5H). In other words, the 'VND KO' correctly informs us that SNA does not position the ventral boundary of *ind* expression. Thus, these three examples show how the information gained by an experiment can be interpreted by examining unique

aspects of predictions of that experiment's filtered ensemble on a second experiment.

2.6 QUANTIFYING AND INTERPRETING THE VALUE OF PERTURBATION EXPERIMENTS ON THE *SIM* ENHANCER

Similar to ind, single minded (sim) is dorso-ventral patterning gene in D. melanogaster that has been the subject of many biological experiments that describe the regulators of the gene, delineate its enhancer [58, 59, 60], and characterize the combinatorial action of multiple TFs and cell signaling in the formation of the precise expression pattern driven by the *sim* enhancer [61]. The *sim* gene is initially expressed at the cellular blastoderm stage in a narrow row of width equal to two cells along the dorso-ventral axis at the mesectoderm (the boundary between mesoderm and neural ectoderm) [59, 60] (Figure 2.6A). *sim* acts as a master regulator during the development of central nervous system (CNS) [62] and the confinement of its expression to the narrow line of cells is essential for the formation of the ventral midline and CNS during gastrulation [58, 63]. This precise pattern of expression can be explained by a complex regulatory mechanism that involves Notch signaling [64, 65, 66]. On the ventral side, DL and Twist (TWI) activate but SNA represses the expression in the mesoderm [58, 63]. Expression on the dorsal side is inhibited directly by Suppressor of hairless (Su(H)), which is the only known repressor of *sim* in the neuroectoderm [64], but is believed to have an activating influence on *sim* in the mesoderm region [56, 64, 67, 68]. With these pieces of mechanistic information in hand, we employed GEMSTAT to model the expression driven by the *sim* enhancer. Then, we used the procedure introduced above to examine different perturbation experiments related to this enhancer reported in the literature, and quantify and interpret the 'value' of these experiments, after the fact. To our knowledge, this work is the first attempt to computationally model the expression of the *sim* enhancer, using the combinatorial action of TFs and signaling [64, 66, 69].

We built an ensemble of models that predicts the wild-type expression profile of *sim* accurately (Figure 2.6B and Methods) from its wild-type enhancer ('2.8sim'). We then considered several experiments reported in the literature pertaining to this gene, with the goal of computing the information gain from each experiment and interpreting the information they provide. Each of the nine experiments considered is a reporter assay with a variant of the wild-type *sim* enhancer, and we used its observed readout to construct objective criteria (S2 Table) for filtering models and creating a 'filtered ensemble' for that experiment (S6 Fig). This allowed us to quantify the information gain score of each experiment, using the procedure described in previous sections (Figure 2.7A, B). This revealed that the experiment '$2.8sim\Delta SD16$', representing a deletion of two segments (harboring a SNA site and

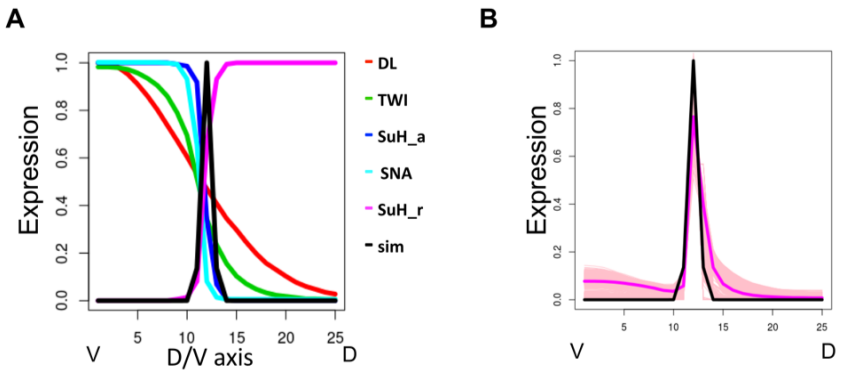

Figure 2.6: (A) Modeling *sim* enhancer. Expression profile of *sim* and all TFs that are involved in *sim* regulation, shown for ventral-most bins 1-25 of the 50 bins along D/V axis. Su(H) is modeled as both an activator and a repressor. (B) Ensemble of models that predict *sim* expression profile accurately.

an E-box element respectively) from the wild-type enhancer 2.8sim, is the most informative (value 2.25), while seven of the other eight experiments are substantially less informative (about 0.5 or less). Following the procedure of the previous section, we then sought to interpret the information gained by this experiment. This was most apparent when we used the filtered ensemble of this experiment to predict the outcome of another experiment ('mesectoderm2.2').

This second experiment is the reporter readout of a 2.2-kb sequence upstream of the early *sim* promoter and overlapping with the wild type enhancer 2.8sim considered above . According to the literature [69], the expression driven by this sequence is unchanged (Figure 2.7C) from wild-type. The filtered ensemble of the '2.8$sim\Delta SD16$' experiment [58] can predict this known outcome accurately (2.7D), while the wild-type ensemble is far more uncertain in its prediction (2.7E). The '2.8$sim\Delta SD16$' experiment tests the effect of deletion of SNA sites on the gene expression and restricting the wild-type ensemble based on results of this experiment informs us which group of SNA sites is important to set up the precise expression of the *sim* gene.

2.7 DETAILED DESCRIPTION OF THE ENSEMBLE ANALYSIS

2.7.1 Construction of Model Ensemble

We used the GEMSTAT model from [36] with 13 different parameters, including two parameters for each of the five TFs: one pertaining to TF-DNA binding (K_{TF}) and one to the TF's effect on transcription rate (α_{TF}). Moreover, the model has one parameter for DL-ZLD

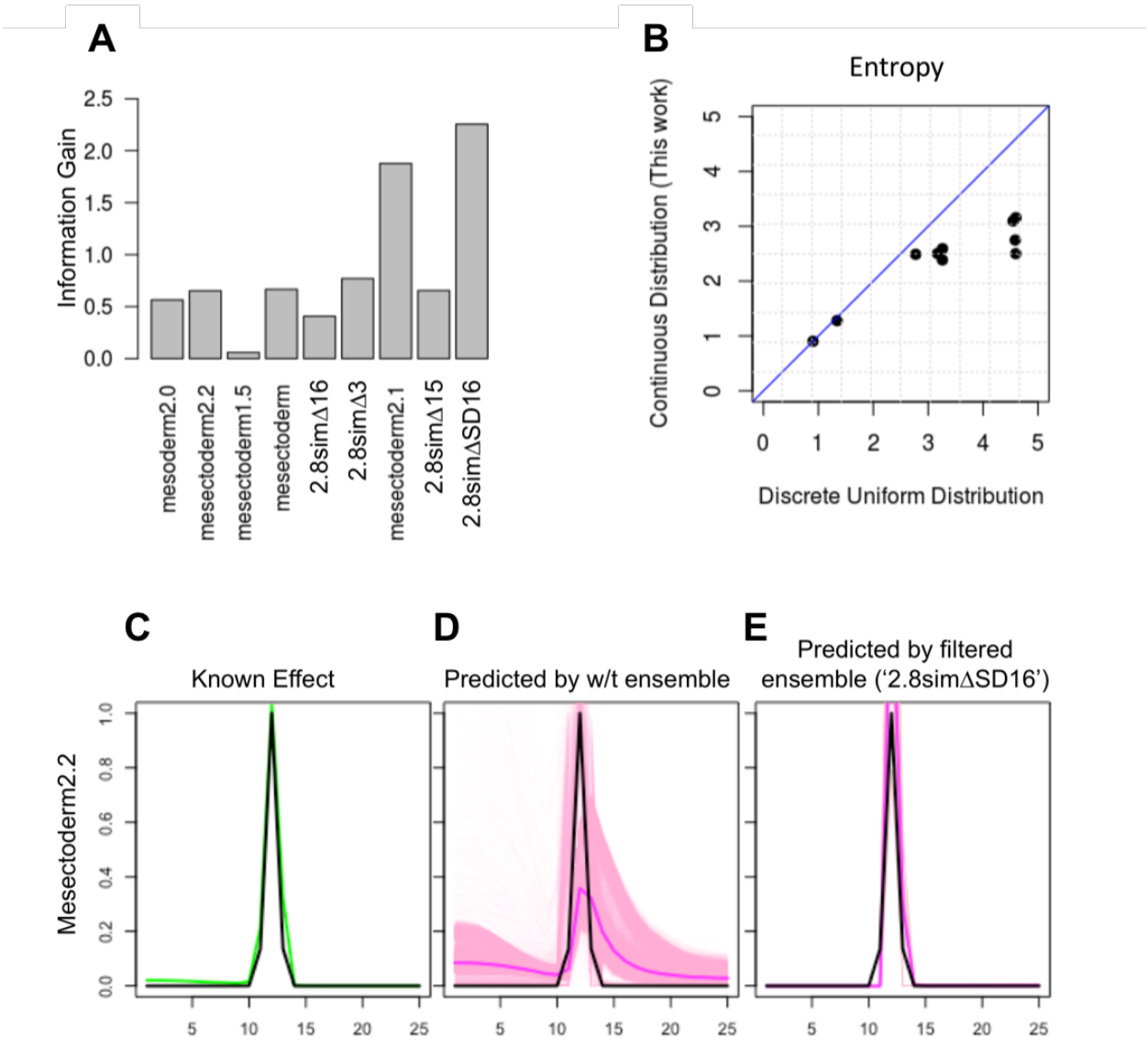

Figure 2.7: (A) Information Gain from different perturbation experiments. Each experiment represents the readout of a variant of the wild-type enhancer, under wild-type conditions, and is named for the variant enhancer. (B) Entropy of filtered ensemble for each of the nine experiments, as defined under the specially constructed probability distribution presented in this work (Y axis) or under a discrete uniform distribution (X axis). (C-E) The 'Mesectoderm2.2' variant of the *sim* enhancer [69] (S2 Table) has been observed to recapitulate the expression pattern of the wild-type enhancer (D). In contrast, majority of the models in the wild-type ensemble (E) predict expansion of the dorsal boundary, but an ensemble filtered by the '2.8$sim\Delta SD16$' experiment predicts the known (non-)effect correctly.

cooperativity (ω_{Dl-Zld}), and another parameter reflecting the baseline transcriptional rate (q_{BTM}). The repressor CIC has a uniform dorso-ventral expression profile but its repressive effect is attenuated in the neuroectodermal region by locally activated ERK, through a reduction in CIC-DNA binding; this effect, as modeled in [20], is represented by a free parameter (Cic_{att}). The expression pattern of each TF and the *ind* gene was scaled in the range of zero to one. Each expression profile is represented by a 50-dimensional vector (S1A Fig), with dimensions corresponding to equally spaced bins along the D/V axis (bin 1 = ventral end). The *ind* gene is expressed in only 5 to 7 of these bins (bins 22-28 with the peak of expression at bin 25).

To construct the ensemble, we sampled models from the 13-dimensional parameter space following the procedure of Samee et al. [20]. We divided the range of each parameter into two halves (using log scale for K parameters of all TFs, α parameters of repressors and for q_{BTM} , and linear scale for: α of all activators, ω_{Dl-Zld} and Cic_{att}), and sampled 1000 points in each cell of the 13-dimensional space, for a total of 8 million models. Such a dense sampling was possible because the number of parameters (13) is modest. Among these randomly generated models, we retained those with SSE (sum of squared errors) score between the real and predicted *ind* profile less than 10%. This higher initial threshold allows us to get good initial points. We then used GEMSTAT's optimization routine to locally optimize the initial ensemble, following which we filtered for models that have SSE score less than 5%. This stricter error threshold was determined by a visual inspection of many examples, as one that allows the main spatial pattern of expression to be preserved in the predictions meeting that threshold (see S1 Appendix for further details). We call the resulting collection the wild-type ensemble of models. It contains more than 5000 distinct parameter settings from about 600 different regions, and all of these models make good predictions on the wild-type data, as per visual inspection (Figure A.1 B).

2.7.2 Construction of Probability Distribution over Ensemble of Models

Each model is represented by its parameter vector, and we first scaled the parameters using min-max scaling to place all the dimensions on the same scale of 0-1. The models were then clustered based on a multivariate density estimation method, Mclust [70, 71], using its R implementation. This method approximates the complete collection of models as a (generally non-uniform) mixture of Gaussian distributions, each component Gaussian representing a cluster, with its own mean (cluster center) and covariance matrix (S5 Fig), while simultaneously determining the optimal number of clusters. Next, we modeled each cluster C as a uniform mixture of $n_C = |C|$ Gaussian distributions with each of the n_C models

of the cluster as mean, and a common covariance matrix Σ_C estimated for the cluster by the Mclust method. The following formula describes the probability distribution over the space of models θ:

$$P(\theta) = \frac{1}{N} \sum_{C=1}^{N} \frac{1}{n_C} \sum_{i=1}^{n_C} \mathcal{N}(\theta; \mu_{iC}, \Sigma_C) \tag{2.1}$$

Here N is the number of clusters, C indexes these clusters, n_C is the number of models in cluster C, μ_{iC} is the i'th model in the cluster C, and Σ_C is the covariance matrix of cluster C.

Since models in the ensemble are not simply random samples of the probability distribution, but a collection of local optima obtained from initial random samples as seeds, we did not use a standard density estimation technique to build a density function. (See Discussion for our choice of the above methodology as opposed to more sophisticated sampling techniques.) We expected the distribution to reflect the fact that each model is a local optimum in the landscape of models scored by SSE measure and they group into several clusters. These clusters reflect different hypotheses for the biological mechanism of the underlying system and we desired that the probability density function put equal weights on them. Within each cluster, it is possible to observe several equally good local optima (peaks of the SSE landscape); we selected each such optimum as a local peak for the probability density function as well, constructing a Gaussian Mixture with a fixed covariance matrix to represent a cluster, with component Gaussians centered on the optimized models in that cluster. It is worth noting that our modeling of the desired probability distribution as a mixture of Gaussian distributions has implicit ties to the 'MaxEnt principle', since the Gaussian is the maximum entropy distribution under a given mean vector and variance/covariance matrix.

2.7.3 Entropy of Probability Distribution over Ensemble

We calculated the entropy function using a discrete version of the probability density function in (1). The discrete probability p_{iC} of the model i in the cluster C is set to be proportional to the value of the continuous probability density function at the location of the model. Each p_{iC} receives contributions from all Gaussians in the mixture model. We set the constant of proportionality such that for each cluster C we have $\sum_{i=1}^{n_C} p_{iC} = \frac{1}{N}$ where N is the number of clusters. We then estimated the entropy of this discrete probability density function using the formula $H(P) = -\sum_{C=1}^{M} \sum_{i=1}^{n_C} p_{iC} \log p_{iC})$ [72].

2.7.4 Information Gain from an Experiment

Suppose we are given an ensemble of models obtained from a set of experiments and have calculated its probability distribution. We are then given the results of a new experiment. This is typically in the form of gene expression level(s) driven by an enhancer sequence in one or more cellular contexts described by their TF concentration profiles. We assess if the predictions made by any model in the ensemble are consistent with these results, by comparing the model's predictions to the observed gene expression levels through a single goodness-of-fit score such as SSE (sum of squared error), and discard that model if this score is worse than a pre-determined threshold. Repeating this process for each model in the ensemble, we obtain a 'filtered ensemble' that is a subset of the original ensemble. We then modify the probability distribution over the filtered ensemble by (a) retaining the cluster assignments and cluster covariance matrices of the original ensemble, but (b) removing any cluster that has no remaining models and readjusting weights of remaining clusters to add to 1, and (c) redefining the Gaussian Mixture Model for each cluster to have component Gaussian distributions centered at each model remaining in that cluster (also with uniform weights). Having thus defined a probability distribution over the filtered ensemble, we calculate its entropy as above and assign the difference of entropy between the (distributions over the) original and filtered ensemble as the information theoretic 'value' of the new experiment.

2.7.5 Data sets for modeling *sim* enhancer

We constructed a one-dimensional vector to describe the expression readout of the *sim* enhancer along the D/V axis, as recorded in the literature. This vector (and other expression vectors described here) has 25 dimensions, representing equally spaced positions ('bins') along the axis from the ventral end to the mid-point of the axis. (These are the same as bins 1-25 of the 50 bins considered in previous sections.) The expression in each bin has a value between 0 and 1 that corresponds to the relative amount of gene expression observed in that bin. The expression profile of the wild-type *sim* enhancer was represented as a Gaussian curve that has its peak at the location where SNA expression changes from high to low (12th bin from ventral end), i.e., at the known location of the *sim* expression peak. The variance of the Gaussian is set to be small enough that the 'width' of the expression profile is similar to the narrow domain in which SNA goes from high to low (2.6A). We obtained TF (protein) expression profiles of DL, TWI, and SNA from Zinzen et al. [32] and represented them in the same 25-dimensional vector format as above (2.6A). We considered

DL and TWI as activators and SNA as a repressor. The other important regulatory input considered was Su(H), which is a maternal protein uniformly expressed across the D/V axis. It is believed to be a repressor, but Notch signaling activated by the effect of SNA on Notch-Delta endocytosis switches the role of Su(H) from a repressor to an activator [64, 65, 66] in domains of SNA expression (mesoderm). Since GEMSTAT does not allow for such a 'role-switch' for any TF, we separated the uniform expression profile of Su(H) into two separate profiles (vectors), one for each role: an 'activator Su(H)' with an expression profile similar to SNA but extended to include the mesectodermal positions and a 'repressor Su(H)' with its complementary profile. In this manner we capture the prior knowledge of the 'role-switch' of Su(H) at the peak expression of sim. The *sim* enhancer sequence and TF motifs, required by GEMSTAT, were taken from Fly Factor Survey [73].

2.8 DISCUSSION

Determining regulatory mechanisms shaping the spatio-temporal pattern of a gene of interest is a tedious process. While high throughput technologies provide helpful clues and narrow the space of possibilities, the 'gold standards' for demonstrating the regulatory influence of a transcription factor on a gene – a combination of TF knockout or overexpression (and observed effects on gene expression), TF-DNA binding assays, site mutagenesis and rescue experiments – involve substantial investments. Guidance about the most insightful experiments to perform, given current knowledge about the gene's regulation, can thus be highly beneficial. Typically, such choices are made by the biologist by relying on their intuition. We asked ourselves if the process of designing experiments to gain deeper understanding of a gene's regulatory mechanisms may be made systematic. This immediately presented two major conceptual challenge: first, how do we formalize what is 'current knowledge' about the gene's regulation, and second, how do we measure how insightful or informative an experiment is? Answers to these questions appear to be necessary before we could systematize the process of experiment selection or design, mentioned above. In this manuscript, we take present a possible solution to these challenging problems by making use of a previously established quantitative modeling framework that relates trans- and cis-regulatory information to gene expression levels, and combining the framework with ensemble modeling and information theoretic ideas. In the future, our approach can be combined with well-established ideas in statistical experiment design [47, 48] to develop a full-fledged formal approach to investigation of gene regulatory mechanisms.

We approached the goal of formalizing current knowledge about a gene's regulation by using the GEMSTAT framework of gene expression modeling. (Other related models, e.g.,

[74], would also have been similarly usable.) Here, current knowledge of a gene's enhancer sequence(s) and its known regulators (TFs) is encoded into a mathematical function that is consistent with data on the gene's and TFs' expression levels in multiple conditions or cell types. This not only forces the qualitative knowledge of regulators into a precise quantitative form, it also explicitly captures complexities and subtleties associated with combinatorial action of multiple TFs. Furthermore, the model has free parameters representing important but often uncharacterized biochemical properties of the regulators, viz., free energy of DNA binding and strength of regulatory influence, and the modeling step involves assigning values to these parameters so as to match available data. In this step, one is often faced with many distinct parameter settings that appear equally plausible in light of available data, and these different parameterizations represent ambiguities in current mechanistic understanding of the gene's regulation, even when the likely regulators are qualitatively characterized. In our approach, such ambiguities are explicitly catalogued in the form of an ensemble of models consistent with data. To address the other conceptual challenge mentioned above ('how informative is an experiment?'), we compared the ensemble representing prior knowledge/data to that representing new experimental data in addition to the prior information. It was natural to consider using the information theoretic ideas of entropy and information gain for purposes of this comparison. We therefore devised an approach to define a probability distribution over models in the ensemble, and to estimate the entropy of the distribution; the information gain was then defined as the difference in entropy of the two ensembles.

We demonstrated the use of our approach in the context of two genes in early fruitfly development – *ind* and *sim* – whose regulatory mechanisms have been studied through several perturbation experiments (TF knockouts, site mutagenesis, variant enhancers, etc.) reported in the literature. In each case, we started with the wild-type enhancer and likely regulators as 'current knowledge' and (retroactively) quantified how informative each of the perturbation experiments is. We also presented objective observations about each experiment that suggest the specific insights it added to our understanding of the gene's regulatory mechanisms. In the case of ind, we additionally applied our experiment-scoring framework in a more controlled, semi-synthetic setting, where real data on the gene were used to first select a unique model as the underlying 'truth', and used to provide the results of in silico perturbation experiments. We note however that the information gain values computed by our method are not comparable across different studies, e.g., between the *ind* and *sim* studies considered here; they are only comparable across different experiments for the same gene, when evaluating those experiments for additional insights over a common set of current data/knowledge.

Methodologically, an important feature of our approach was the generation of ensemble by

uniform sampling in the multi-dimensional space, followed by optimization, as was done in [20]. Alternative sampling algorithms can be used to generate a large sample from optimal regions of the parameter space. Sampling methods such as Bayesian optimization techniques can be used to efficiently search for optimal parameters of any given model with nonlinear cost functions [75]. However, in practice these sampled optimal solutions may not be representative of every possible region of the parameter space with similar goodness of fit. One example of such a Bayesian optimization algorithm is the spearmint package. Spearmint, when applied to our problem, produced only a few optimal models (data not shown). Building a large ensemble of models that represents every locally optimal region requires running the method multiple times, which is very slow, due to slow convergence time [76]. The spearmint method yields only few data points as the result of optimization and they are usually close to each other. Unless we perform a detailed sampling around those points or combine such techniques with a more global sampling approach, we do not have a diverse ensemble to work with. On the other hand, since the number of parameters in our model is small (less than 20), we could afford to do a dense uniform sampling of the parameter space with our approach. Our ensemble generation process has ties to Bayesian inference, as it constructs a distribution over models M, given data D. The more common approach to this is to sample from the posterior distribution P(M—D), using a suitable likelihood model. There are two reasons why we chose not to do this in our approach. First, we wished to impose upon the distribution the property that different high-density regions of the space have equal relative weights (probability mass), since each of these regions represents a distinct mechanistic hypothesis to us. This property is technically challenging to encode in the form of a prior distribution, and would require substantial research into the Bayesian inference methodology, which was not our main focus. Second, we found that a standard Bayesian optimization technique (which we tested) was not very efficient at sampling the parameter space globally.

We believe that the framework established in this work can be used in future work to formalize experiment design strategies for gene regulation studies. For instance, we may compute the expected information gain [77] of various possible future experiments and select the best one. Given a candidate future experiment, we may first use each in the current ensemble to predict its outcome, compute the resulting information gain, and then compute an expectation of this value over the entire ensemble, using the probability distribution introduced in our work.

CHAPTER 3: MODEL-BASED ANALYSIS OF POLYMORPHISMS IN AN ENHANCER REVEALS CIS-REGULATORY MECHANISMS

3.1 UNDERSTANDING THE GENE REGULATORY MECHANISMS USING ENSEMBLE OF MODELS

Sequence-to-expression models have been proposed in the literature to address the above need [78]. These are mathematical models, based on biophysical principles [74] or machine learning concepts [28, 27], that map an enhancer's sequence, optionally along with additional contextual information such as cellular concentrations and DNA-binding preferences of TFs, to the expression level driven by that enhancer. These models formalize what is known about a gene's regulatory mechanisms encoded in enhancers, and have proven capable, in some cases, of predicting the effects of minor sequence differences such as mutagenesis of entire binding sites [18, 20, 21]. However, when using these models one is faced with a trade-off in predictive accuracy: one fits the model to many different enhancers [36] if one wishes to capture broad regulatory mechanisms, but the resulting models are not capable of predicting the effect of minor changes such as single nucleotide variations. To achieve this latter capability, one typically fits the model to fewer, more closely related enhancers [20], but this results in under-constrained models and parameter uncertainty [42]. The result is not a unique trained model but an ensemble of models, representing distinct mechanistic explanations of the data, and thus an ensemble of predictions about the effect of the same sequence mutation. If additional information becomes available about the true effect of an enhancer mutation on gene expression, that information may be found to be consistent with only a subset of the current ensemble of models and thus allow us to filter the ensemble and reduce our uncertainty about parameters, consequently increasing our confidence about cis-regulatory encoding of the enhancer. This is the key insight we pursue in this work.

We first used a thermodynamics-Based modeling framework to fit an ensemble of models that relate the expression pattern of the gene intermediate neurons defective (ind) to the known enhancer of the gene. Thermodynamics based models are among the most successful genre of quantitative models for the sequence-to-expression relationship [18, 74], and formalize the enhancer's cis-regulatory encoding through model parameters that are fit to the available data. We had previously shown how ensemble modeling of the thermodynamics-based GEMSTAT model [20, 42] can provide useful insights about the ind enhancer. Here, we used the ensemble of models to predict the effect of each single nucleotide mutation in the enhancer, and used our previously published probabilistic framework to identify mutations with high expected impact and/or high variance in predicted impact. We experimentally

tested the effect of such mutations, using transgenic reporter assays in fruitflies, and used the resulting additional information to reduce the ensemble of models to a single tightly clustered set of models that represent a unique mechanistic explanation of the enhancer's function. We then showed that the resulting model is indeed supported by additional data not used in the modeling, e.g., it provides better fits to unseen enhancers related to the ind enhancer. Our work attempts for forge a path forward towards deeper mechanistic understanding of the cis-regulatory 'code' [79] and its use in predicting the impact of single nucleotide variants [80].

3.2 THERMODYNAMICS-BASED MODELING OF GENE REGULATORY MECHANISMS

We set up the GEMSTAT model with 13 different parameters as in [20, 42] and trained the model to relate the wild-type expression profile of the ind enhancer to the concentration profiles of the five TFs DL, ZLD, SNA, VND, CIC. GEMSTAT uses the following parameters: K_TF parameter indicating TF-DNA binding (one for each TF), α_{TF} parameter to capture the TF's effect on transcription rate (one for each TF), and the parameter q_{BTM} for the basal transcriptional rate. Based on previous experimental studies [54], DL and ZLD work cooperatively and this was modeled through the cooperativity parameter ω_{DL-Zld}. Previously reported reduction in CIC-DNA binding by locally activated ERK is modeled using the Cic_{att} parameter, as explained in [20]. Henceforth, we refer to any setting of values for the above 13 parameters. We uniformly sampled the defined range of each parameter and measured the SSE score between the wild-type and predicted expression profiles. We used a loose threshold to filter for models with good fits (SSE$<$ 0.15), used these models as starting points for optimization (following GEMSTAT's in-built optimization) and then selected optimized models that meet a strict threshold (SSE$<$ 0.05). The result is an ensemble of 5237 models with distinct parameter settings that produce high quality fits to the expression profile of the gene. This is called the "wild-type ensemble", as it was trained solely on the wild-type expression profile.

Additionally, we utilized data from six different perturbation experiments pertaining to ind gene expression in the same developmental stage as above. These biological perturbation experiments included (i) 'DL 1 site' [55], where mutagenesis of the strongest DL site results in no significant change of ind expression, (ii) 'DL 3 sites' [53], where mutagenesis of three DL sites results in greatly diminished ind expression, (iii) 'ZLD sites' [20], where ind peak expression reduces to 50% of wild-type levels upon mutagenesis of four ZLD sites, (iv) 'SNA KO' [32], where knock-out of SNA results in no significant change, (v) 'VND KO' [56],

where knock-out of VND leads to ventral expansion of ind expression, and (vi) 'CIC site mut.' [57], where CIC site mutagenesis leads to dorsal de-repression. Each of the 5237 models in the wild-type ensemble was used to predict the effect of each of these six perturbation experiments, and only those models whose predictions were consistent with data for at least 5 out of the 6 perturbations were retained (Supplementary figure A.13). Only 12 of the 5237 examined models met this requirement and none of these makes predictions consistent with all six perturbation experiments; in particular, no model was able to explain the 'DL 3 site' and 'ZLD sites' perturbations simultaneously. To obtain a larger ensemble of models similar to these twelve, we sampled 10000 points in the parameter space around each of the 12 models and optimized the parameters to fit the wild-type expression profile of ind, using the sampled points for initialization. We retained from among the resulting models only those whose predictions were consistent with the perturbation experiments by the above-mentioned criterion. The retained models cluster into 18 different groups (as per method noted in the next paragraph), and we sampled 100 models from each group to obtain an ensemble of 1800 models in total. This collection of models is referred to as the "filtered ensemble". Systematic prediction of SNP effects using ensemble of models

We then followed the procedure in our previous work [42] to first cluster all models in an ensemble and then construct a probability distribution over the models such that each cluster (or group) of models has the same overall probability. We then computed the average predicted effect of every possible single nucleotide mutations in the ind enhancer. The effect of a mutation was computed by comparing a model's predicted expression profile for the wild-type enhancer to that for a mutated enhancer (carrying the specific mutation), and recording the sum-of-squared-errors (SSE) between the two profiles. We repeated this procedure for every model in the ensemble and computed the mean (as well as variance) of predicted effects, over the above-mentioned probability distribution over the ensemble. Such ensemble averages were separately computed for the wild-type ensemble, the filtered ensemble as well as for each cluster of models within the latter.

3.3 STATISTICAL TESTING OF COMPENSATORY EFFECTS OF SNPS

We generated 2000 synthetic genotypes representing the ind enhancer that exhibit polymorphisms at the 70 SNP positions in the DGRP population, at the same allele frequencies as in the population. For this, we first represented the DGRP genotypes as a genotype matrix where the rows represent SNPs, columns represent lines and each cell in the matrix is 0 or 1 depending if a line carries the SNP. We permuted this matrix while preserving the sum of each row as well as each column. The permutation process selects at random two

rows and two columns such that each row and each column of the resulting 2 x 2 matrix has exactly one '0' and one '1', and swaps its rows, and the repeats this operation 1000 times. At the end, a column is chosen at random from the resulting permuted matrix and represents a sampled genotype. For each sampled genotype, we made model-based predictions of the effect (SSE) of all mutations present in that genotype as well as the effects of each mutation individually, and deemed the genotype as exhibiting compensatory mutation if the effect of all mutations together was less than the strongest among individual mutation effects. We compared the number of genotypes with compensatory mutation (964 of 2000) to the corresponding number in the DGRP (168 of 205), using a Fisher's exact test.

In vivo reporter assays

ind1.4 enhancer constructs were prepared using as a template the wild-type version of the enhancer present in DGRP line RAL-821. Mutagenized enhancers carrying changes at positions 1198, 309 and 309 + 324 ("construct1", "construct2", "construct3" respectively, see Results) were generated by recombinant PCR and sequenced to confirm their integrity. The final transgenes were assembled in the placZ-attB vector [81] and integrated at chromosomal position 86F via C31-mediated germ-line transformation [82].

3.3.1 Model Predictions for *rho* Enhancer

These models had eight parameters, four of which ($K_{DL}, \alpha_{DL}, K_{SNA}, \alpha_{SNA}$) were shared with models for the ind enhancer and were kept fixed at values trained on the ind data, while the other four (K_{TWI} , α_{TWI}, the cooperativity parameter ω_{DL-TWI} , and q_{BTM}) were trained on *rho* data set of Sayal et al. [18].

3.4 EXPRESSION DATA SUPPORT DIVERSE MECHANISTIC MODELS OF *IND* ENHANCER FUNCTION

Our first goal was to train a model capable of predicting the impact of enhancer sequence changes on gene expression. For this, we considered various models in the literature that can predict gene expression profile from an enhancer's sequence and information about transcription factor (TF) concentrations and their DNA binding preferences (motifs) [35, 31, 32, 18]. We chose to work with one such model, called GEMSTAT [36], which was previously reported by us and successfully used to model several developmental enhancers of Drosophila [20, 49]. GEMSTAT uses a statistical thermodynamics formulation to capture the molecular interactions between TFs and DNA and their quantitative impact on transcription rate. It uses two tunable biophysical parameters for each TF: one parameter that represents acti-

vation/repression strength and the other related to DNA-binding strength of the TF at its optimal site. It has one additional global parameter corresponding to the basal transcriptional machinery and optional parameters for cooperativity between specific pairs of TFs. GEMSTAT uses available data – enhancer sequence(s), expression levels, TF concentrations and TF binding specificities or 'motifs' – to find optimal values for its free parameters (usually 10-20 parameters, representing 5-10 TFs), and in some cases this procedure is known to result in locally but not globally optimal parameter values. We addressed this problem in recent work [42] by generating and reasoning with an ensemble of model parameters that fit the data, rather than determining a single parameter setting that maximizes the 'goodness-of-fit'.

In the current work, we first modeled the expression data available for the 'intermediate neuroblasts defective' (ind) gene in Drosophila melanogaster, following our previous work [42]. The enhancer for ind is well characterized, and its regulators are well-studied. There are two activators: Dorsal (Dl) and Zelda (Zld) and three repressors: Snail (Sna), Ventral neuroblasts defective (Vnd) and Capicua (Cic). The expression data includes the spatial pattern of ind gene and its TFs in the blastoderm stage of embryonic development, as a 1-dimensional profile along the dorso-ventral (D/V) axis (Figure 3.1 A). By optimizing parameters of the GEMSTAT model through a comprehensive grid-search, we obtained an ensemble ('wild-type ensemble', see Methods) of 5237 models that produce close fits to the wild-type expression pattern (Figure 3.1 B). (Each model is a distinct setting of tunable parameters, see Figure 3.1 C.) We next challenged the ensemble of models with published data [20, 53, 54, 55, 56] on ind gene expression under various perturbation conditions such as mutagenesis of one or more sites of a TF or knockout of a TF. Discarding models in the ensemble whose predictions were inconsistent with results from two or more of these six perturbation experiments, and performing deeper sampling and optimization around the remaining models, we constructed a new ensemble of models, henceforth called the 'filtered ensemble' that reasonably capture ind expression in wild-type as well as perturbation experiments (see Methods). Closer examination of the filtered ensemble and clustering of the parameter vectors revealed multiple groups of models in the ensemble (Figure 3.1 D). The existence of several distinct groups of models is also clear upon visual inspection of a lower-dimensional projection of the parameter vectors using principal components analysis (Figure 3.1 E).

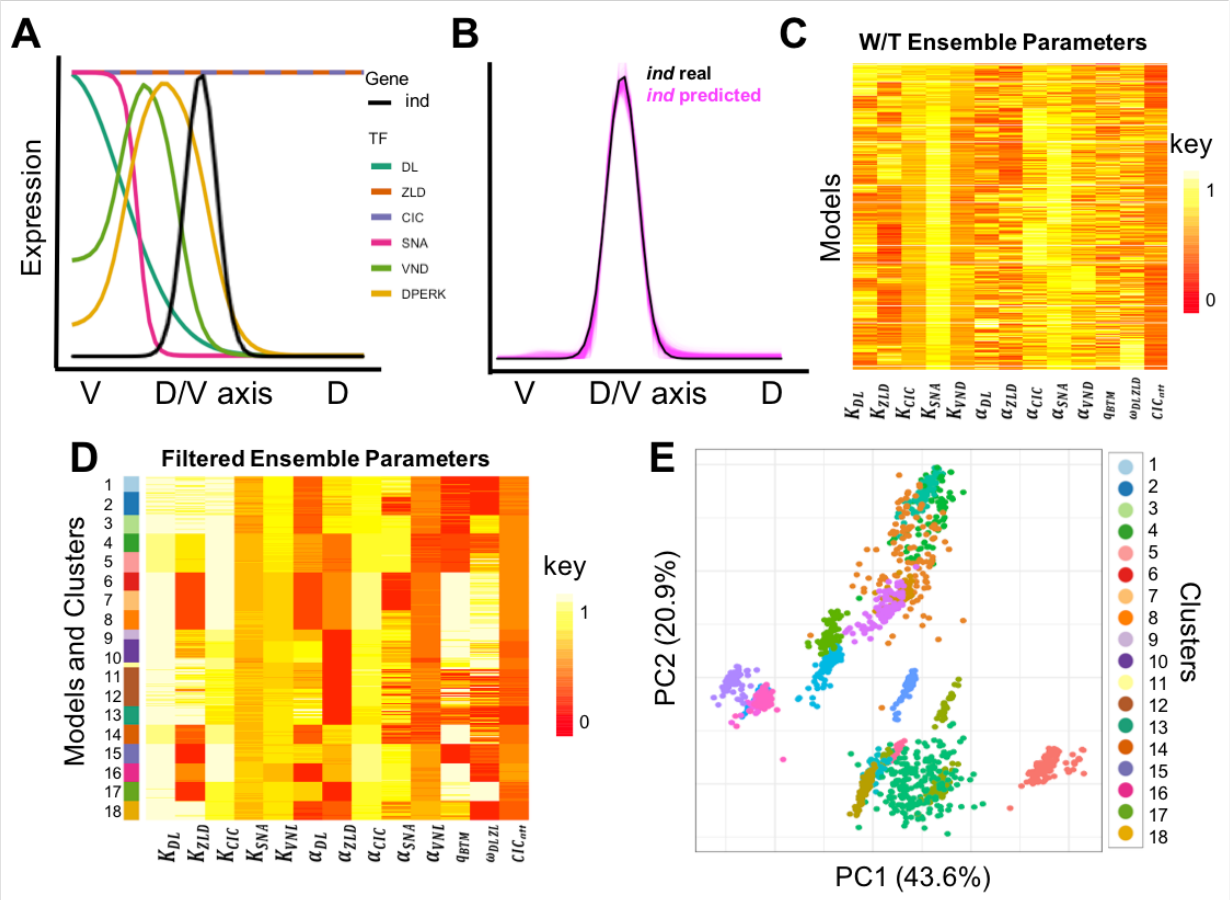

Figure 3.1: Predictions and parameter values of "wild-type ensemble" and "filtered ensemble" of models. (A) The expression profiles for TFs and the 'ind' gene are shown along the Dorsal-ventral domain. The x-axis represents ventral (left) to dorsal (right) end of the D/V axis and the y-axis is the expression value from no expression to the maximum observed expression for each gene or TF, on a scale of 0 to 1. (B) Predicted ind expression (magenta) from all models optimized to fit wild-type data (black). Each pink line shows the prediction of a single model in the wild-type ensemble. (C) Parameter values of models shown in B. Each row is a model in the ensemble and each column corresponds to a parameter, with values scaled to the range of 0 to 1 across all models. The K parameter for all TFs and α parameter of repressors are in logarithmic scale and the α parameter of activators, ω_{Dl-Zld} and q_{BTM} are in linear scale. (D) Similar to C, parameter values of "filtered ensemble" of models, with sidebar colors denoting different groups of models that cluster together (see panel E). (E) Two dimensional projection of 13-dimensional parameter space of the filtered ensemble using the first two principal components. Colors are similar to sidebar in D.

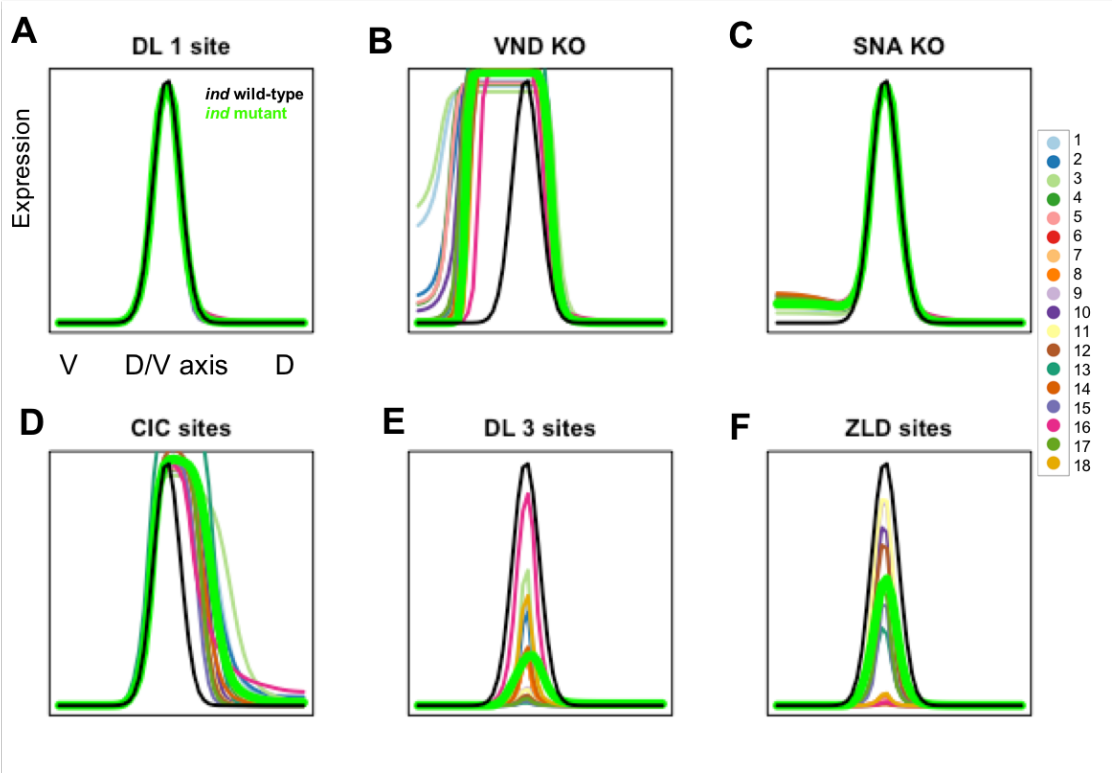

Figure 3.2: (A-F) Predictions of ind expression profile, made by models in filtered ensemble, under six different perturbation conditions, compared to experimentally determined profile under those conditions (green line) and wild-type ind expression (black line). Predicted expression profiles are colored according to the cluster that the predicting model belongs to, with colors being the same as those used in sidebar of panel D in figure 3.1. Experimental data from the literature indicate that ind expression does not change when the strongest predicted DL site in the enhancer is mutated (A), expands ventrally upon VND knockout (B), but is not changed upon SNA knockout (C), and expands dorsally when two binding sites of CIC are mutated (D). Also, peak ind expression is reduced by 65% when 3 DL sites in the enhancer are mutated (E) and peak ind expression is reduced by half upon mutagenesis of ZLD binding sites (F).

The predictions of each group of models in each of the perturbation conditions are shown in Figures 3.2A-F, along with the true expression profiles in those conditions. We noted that all groups of models were consistent with the four perturbation experiments represented by Figures 3.2 A-F, and were additionally consistent with the experiments represented by either Figure 3.2E or Figure 3.2F. Figure 1E reveals uncertainty about mechanisms underlying the gene's regulation, even after subjecting the model to data from the several experimental conditions noted above. Each group or cluster of models represents a distinct hypothesis about the regulatory mechanisms underlying ind regulation and further information is necessary to narrow down the possible mechanisms. We looked to polymorphism data from a population

of D. melanogaster lines [83] for such information, as described next.

3.5 MODEL-BASED ANALYSIS OF POLYMORPHISMS IN THE *IND* ENHANCER

We assessed all possible single nucleotide mutations in the ind enhancer (length 1416 bp, Figure 3.3 A) using the filtered ensemble as follows: for every position in the enhancer, for every possible mutation at that position, we predicted the effect of the specific mutation using each model in the ensemble. We measured the magnitude of the predicted effect as the 'sum of the squared errors' (SSE) between model-predicted expression profile of the wild-type enhancer and predicted profile of the wild-type enhancer modified by that particular mutation. We then summarized the effect (SSE) of the mutation as predicted by all models in the ensemble, using a probability distribution over the filtered ensemble constructed as in [42].

The predicted effect of a mutation is defined as the SSE averaged over the ensemble. Figure 2B shows these predicted effects for every position of the ind enhancer, aligned with the position and strength of each TF binding site in the wild-type sequence (Figure 3.3A). (For each position, only the mutation with greatest predicted effect is shown; see Supplementary Figure A.7 A-B for examples of such effects.) Following the same procedure, we also computed variance of the predicted effect among models in the ensemble (Figure 3.3 C). The 'heatmaps' in Figure 3.3D depict the effect of all possible mutations within three specific transcription factor binding sites (TFBS) that are strong matches to their respective motifs and harbor mutations with relatively large predicted effects. Most models in the filtered ensemble predict that the gene expression should change significantly upon mutating at least one position within these strong binding sites (Figure 3.3D). Predictions were not entirely consistent among models though, and different models vary in predicted extent of change. For instance, Supplementary Figures A.7 A, C reveal discrepancies among the predicted effects of a mutation in a VND site, as predicted by different models in the ensemble. We noted above examples of single nucleotide mutations that could potentially cause a large change in gene expression, although one might expect the more impactful mutations to be avoided in a population, given that our analysis focuses on an early developmental enhancer [84].

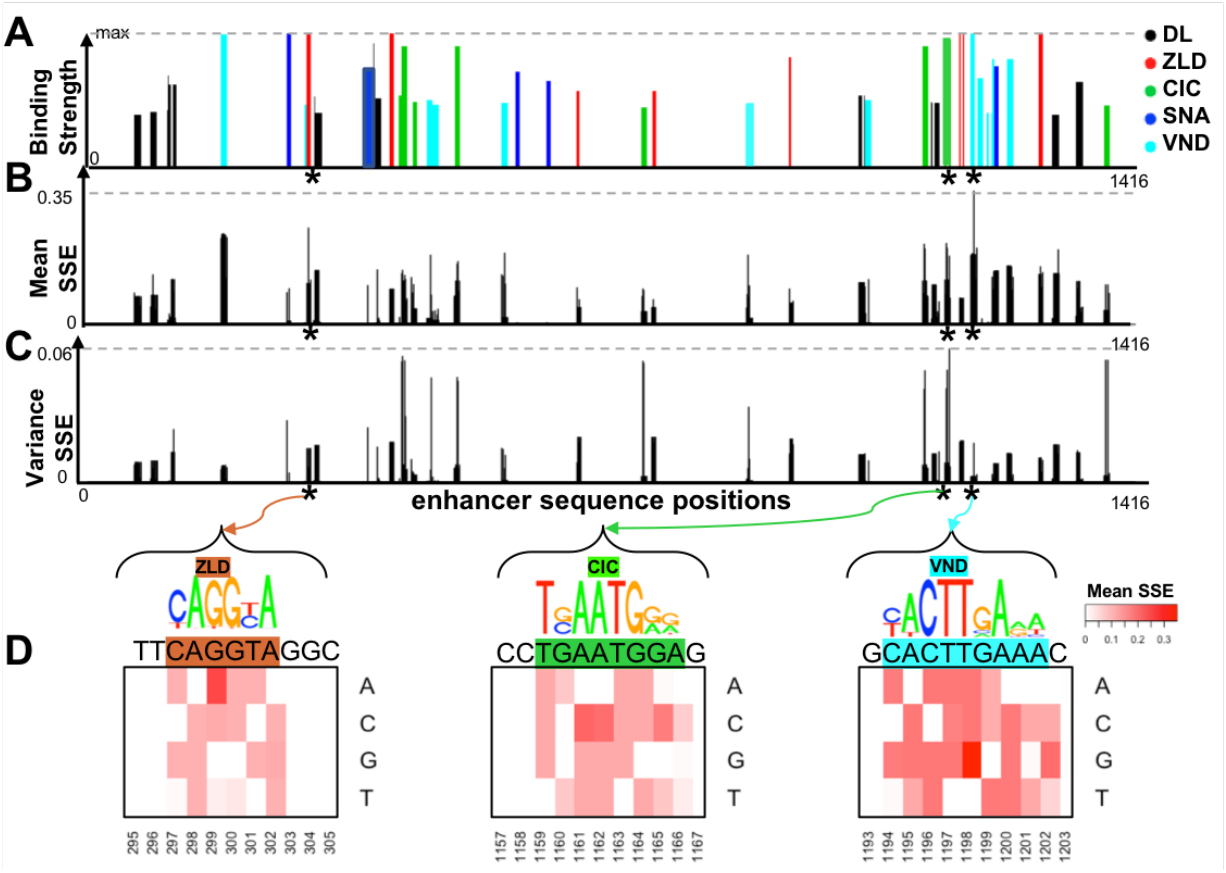

Figure 3.3: Model-based analysis of polymorphisms in the ind enhancer. (A) TF binding sites in the ind enhancer. The x-axis shows positions in the enhancer (base-pairs) and the y-axis shows the predicted binding strength of each TF. (B) Mean predicted effect of each possible mutation in ind enhancer. At each position of the enhancer, we introduce a single mutation (one of three possible changes) and compute the SSE score between the wild-type ind expression profile and the mutant expression predicted by a model. We average the predicted SSE score across all models in the ensemble and show the maximum SSE score among the three possible changes. (C) Similar to part B, but in place of average SSE over models, y-axis shows standard deviation of the SSE score across models. As in B, only the maximum over the three possible changes of the base pair at each position is shown. Stars indicate locations of three binding sites for which predicted effects are examined in detail in the panel below. (D) Heatmaps show the mean SSE score for each possible mutation located inside three selected binding sites. (We selected three enhancer positions with the highest mean SSE and located within binding sites of three different TFs; their encompassing sites are selected for this display.)

We compared these impactful mutations to the allele frequencies of SNPs recorded in a population of 205 lines, as per the Drosophila Genome Reference Panel (DGRP), and noted that these mutations are indeed absent from this population. To further explore the relationship between predicted functional impact and allele frequencies, we examined the 70

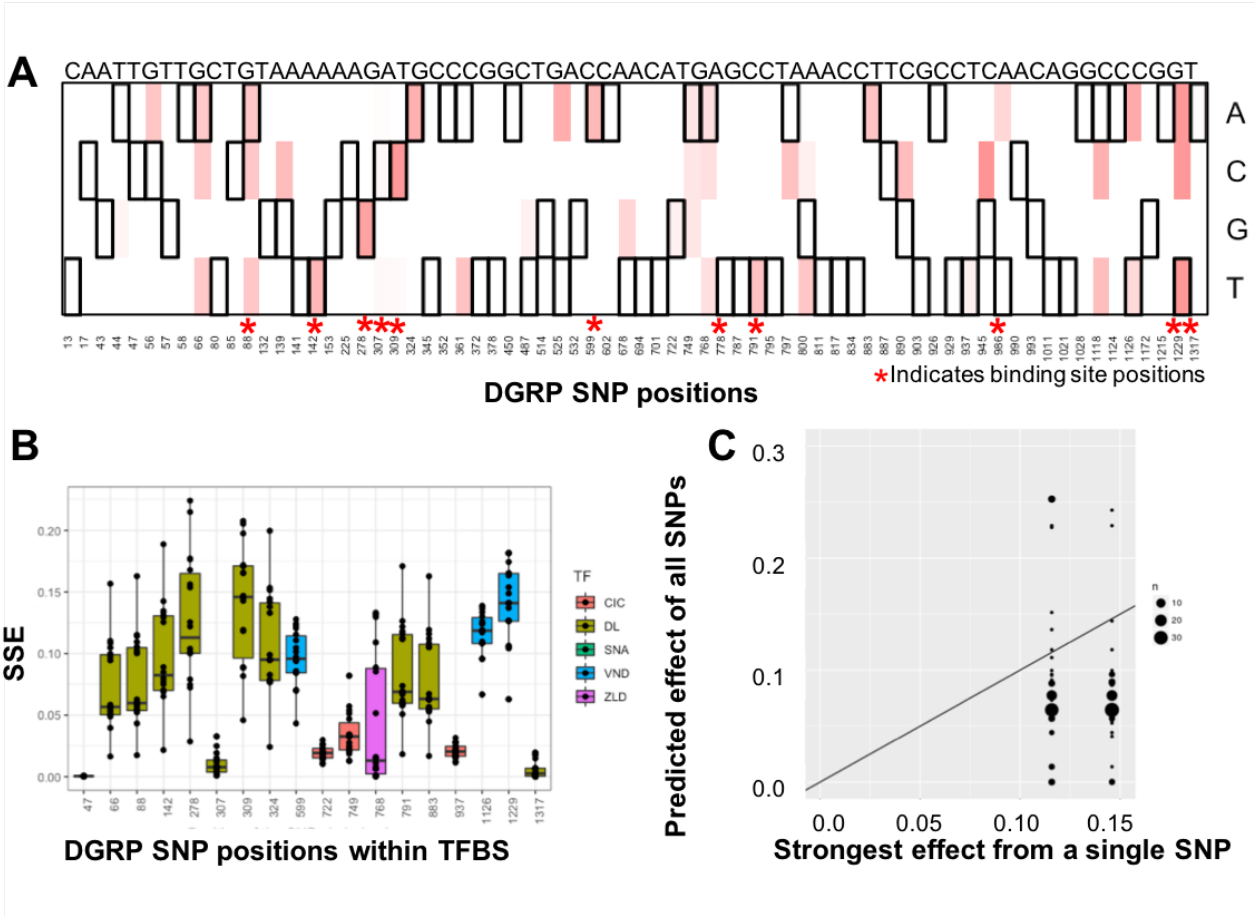

Figure 3.4: (A) Mean SSE scores representing predicted effects of polymorphisms in DGRP. The reference allele is denoted by the sequence at the top and the observed alternative allele is indicated by a black-bordered cell. Red asterisks indicate positions located within TF binding sites in the wild-type enhancer sequence. Positions outside such sites may also have predicted effects, if the alternative allele results in the creation of a site. (B) Distribution (across all models in ensemble) of SSE scores representing model-predicted effect of a polymorphism, for each DGRP SNP. (Only SNPs where at least one model predicts a non-zero effect are included.) Each boxplot shows one such SNP and color indicates the TF whose binding site harbors the SNP. Each black dot is the average SSE score of one cluster of models (from Figure 3.1 D). (C) Each black dot in the scatter plot represents an individual (strain) in the DGRP, x-axis is the highest predicted effect (SSE) of any single polymorphism in the individual, and the y-axis is the predicted effect of all polymorphisms in the individual, taken together. (Predicted effects shown are the mean across all models in ensemble.) Dots of larger sizes represent genotypes with greater frequency in the DGRP. The black line represents the y=x. The plot suggests that most individual enhancers have their largest-effect SNP compensated by other SNPs within the enhancer.

positions in the 1416 bp-long ind enhancer that are polymorphic in the DGRP population (Figure 3.4A). Eleven of the 70 SNPs fall within the annotated binding sites (marked by '*' in Figure 3.4A), and another 7 SNPs give rise to new weak binding sites. In total, 18 of the 70 SNPs may cause changes to binding site strengths and thus to ind expression (Supplementary Figure A.8 E). We used the filtered ensemble to predict the effect of each of the 18 SNP positions separately (Figure 2F and Supplementary Figure A.7 E), and noted that the maximum of these predicted effects – SSE of 0.15 for position 309 – is relatively small compared to the largest predicted effect (SSE of 0.35 for position 1198, Figure 2D and Supplementary Figures A.7 A-C) among all possible mutations, suggesting that high impact mutations are avoided in the population, as expected.

Figure 3.4B and Supplementary Figure A.7 D also reveal that for several of these 18 SNP positions different models in the ensemble make mutually inconsistent predictions (high versus low effect). Such variance in predicted effects points to ambiguities in underlying mechanisms, and offers candidates for experimental testing: data on the true impact of a SNP with ambiguous effect should help constrain the filtered ensemble further and narrow down the viable groups of models further.

3.5.1 Evidence of Compensatory Mutations in Individual Enhancers

Each line in the DGRP population may manifest zero, one or more of the above 18 SNPs, and the net effect of multiple alternative alleles present in a line may not be the sum of their individual effects, i.e., there may be compensatory effects from multiple mutations in an individual enhancer. To investigate this, we next used the filtered ensemble to predict the expression profile of each line's enhancer-level genotype and compared the effect (SSE between this prediction and wild-type expression) to the largest effect from a single mutation carried by the line (Figure 3.2 G). We noted that a great majority (82 %) of lines exhibited signs of compensatory mutations (points below the y=x line in Figure 3.2 G), and that almost all lines were predicted to have an ind expression profile very similar to the wild-type profile (Supplementary Figure A.8 A). We also generated a large number of synthetic genotypes (see Methods) that harbor zero, one or more of the 18 SNPs while preserving allele frequencies of each SNP, and repeated the above exercise (Supplementary Figure A.8 B-D) to obtain a null distribution of the compensatory effect. The fraction of lines that exhibit compensatory mutations in the real population (82%, as noted above) was found to be significantly higher than that in the empirically estimated null distribution (52%) (p-value ¡ 9.62 *10E-22), substantiating the observation of compensatory mutations within lines.

3.6 EXPERIMENTAL TESTING OF SELECTED POLYMORPHISMS IDENTIFIES A SINGLE MECHANISM

We identified (above) several single-nucleotide mutations, present in the population or otherwise, for which the filtered ensemble predicts a large average effect on ind expression or exhibits a high degree of uncertainty. We used in vivo reporter assays to test the expression pattern driven by three such variants (called "construct1", "construct2", "construct3" repecively and explained belo) of the ind enhancer.

The first experiment ("construct1") was designed to test the single mutation that has the greatest predicted effect, averaged over the ensemble. This mutation (Figure 3.5 A), a T -¿ G change at position 1198 in the enhancer, impacts a crucial residue in a high affinity VND binding site, which might result in ventral de-repression of the enhancer, i.e., in its ventral border expanding. The mutation is not seen in the DGRP population, but was selected for the high average and moderate variance in predicted effect. In particular, while the mean prediction of the ensemble of models was a significant ventral de-repression, a subgroup of models in the ensemble also predicted no change in expression, indicating ambiguity in the ensemble.

The second experiment ("construct2") was designed to test a mutation with a high uncertainty, i.e., large variance among predicted effects from different groups of models. This variant (Figure 3.5 B), an $A \longrightarrow C$ change at position 309 in the enhancer, is predicted to reduce the binding strengths of two overlapping binding sites of the activator DL. The predicted impact of this mutation, which is seen in 6.1% of the DGRP lines, is a 25% reduction in the peak height (maximum expression level) on average, but there are groups of models that predict over 50% reduction and those that predict almost no change in expression. Interestingly, the same group of models (cluster 16, Figures 3.3 A,B) predicted the smallest change for both of these enhancer variants.

The third construct ("construct3") tested harbors two single nucleotide differences from the wild-type, and was predicted unanimously by all models in the ensemble to have no effect (Figure 3.5D). The two variants, which are present together in six DGRP lines, include the $A \longrightarrow C$ change at position 309, introduced in the previous paragraph, that should decrease DL binding strength, and another mutation, a $T \longrightarrow A$ change at position 324, also located within a DL binding site and predicted to increase DL binding strength. Together these two changes are predicted by several groups of models to compensate each other, while at least one group predicts neither to impact expression (see Supplementary Figure A.9). We tested the three above variants of the ind enhancer, as well as the wild-type enhancer, through reporter transgenic embryos and confocal laser scanning microscopy (Figure 3.5E,

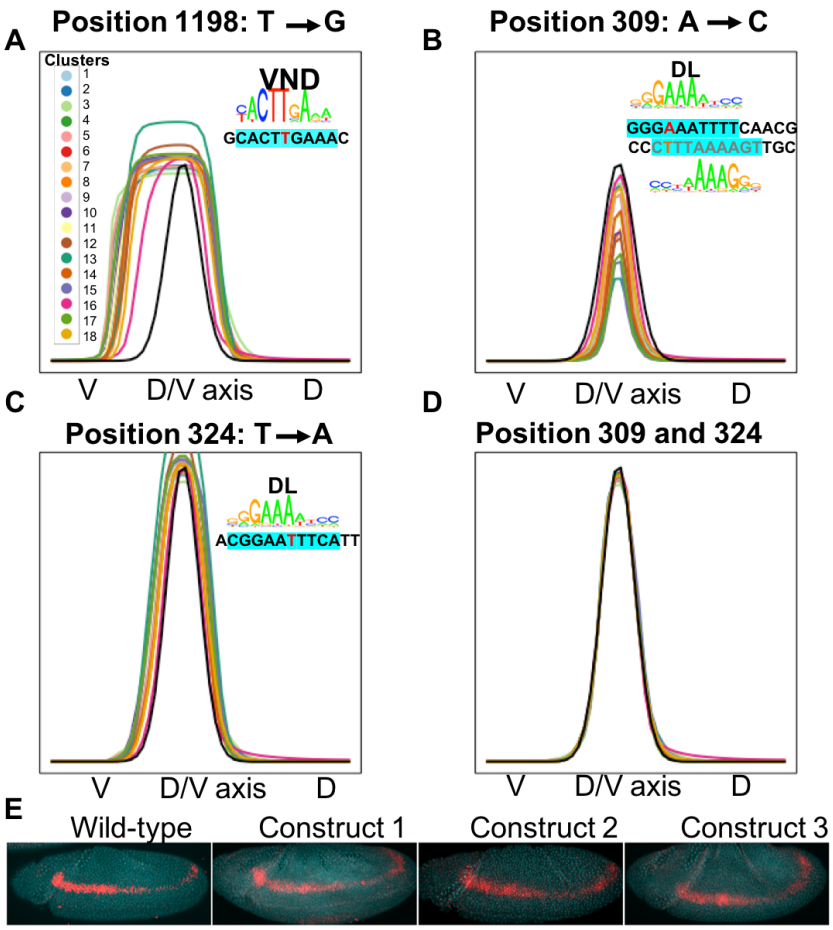

Figure 3.5: Selection and in vivo testing of polymorphisms. (A-D) Model-based predictions of ind expression profile driven by four different variants of the wild-type ind enhancer, each variant differing from the wild-type in either one position (A-C) or two positions (D). Predicted profiles from each cluster of models in the ensemble are averaged and shown separately, in a cluster-specific color. In each panel, inset shows the mutated position in the context of the binding site harboring that position and the motif of the corresponding TF. (A) The selected mutation is located at a key position within a perfect VND site but is not observed in the DGRP population. (B) This DGRP SNP is located within two weak DL binding sites that are conserved across Drosophila species. Multiple groups of models predict a large impact for the mutations shown in A and B. (C) The mutation shown affects a position located within a DL binding site, but the position has a "T' in the wild-type enhancer in place of the preferred "A" (according to the motif); most models predict a modest increase of expression due to a T-¿A mutation at this position, which results from the increased (predicted) binding site strength. (D) Model-based predictions of an enhancer carrying mutations at positions 309 (panel B) as well as position 324 (panel C) of the enhancer suggest that these two mutations have compensatory effects. This pair of mutations is seen in a subset of DGRP individuals (strains). (E) Embryos from flies expressing lacZ under the control of either WT or variant enhancers (carrying mutations shown in panels A,B or D) stained with probes for lacZ mRNA (red) and DAPI (cyan).

see Methods for details). All three variant enhancers were found to recapitulate the wild-type expression pattern of ind. There is exactly one group of models among 18 distinct groups in the filtered ensemble whose predictions are consistent with these new experimental data (Supplementary figure A.10 A-I). In other words, tests of three carefully chosen variant enhancers allowed us to dramatically reduce the space of mechanistic explanations (see Figure 3.1E) to that represented by a tightly clustered group of models, ostensibly representing a single mechanistic explanation of ind regulation. We refer to this group of models as the "final ensemble".

3.7 FINAL ENSEMBLE PREDICTS THE GENE EXPRESSION IN ORTHOLOGOUS ENHANCERS AND VARIANTS OF *RHOMBOID* ENHANCER

3.7.1 Final Ensemble is Consistent with Orthologous Enhancers

Orthologs of the D. melanogaster ind enhancer from other Drosophila species are expected to drive similar expression patterns, given the key role played by this gene in early embryonic development. Under this assumption (also made elsewhere, e.g., [85, 86, 87]), orthologs provide an opportunity to cross-validate models of enhancer function: accurate models when applied to an ortholog may be reasonably expected to predict an expression pattern similar to the known D. melanogaster pattern. We therefore predicted the expression pattern driven by 10 different orthologs of the D. melanogaster enhancer, using the final ensemble alone or using every group of models in the filtered ensemble (Figure 3.5). We noted that the final ensemble makes accurate predictions for the majority of orthologs (Figure 3.6 B, D, F, H, L, N), and provides more accurate predictions on the entire collection of orthologs, compared to other groups of models (Figure 4U. For instance, for the most diverged ortholog – that from D. mojavensis – the final ensemble is the only group of models that predicts expression in the correct location (Figure 3.6 S, T).

3.7.2 Final Ensemble Makes Accurate Predictions on Variants of *rhomboid* Enhancer

In another attempt to test if the final ensemble is more accurate compared to other groups of models in the filtered ensemble, we compared its predictions on the wild-type enhancer of a different neuroectodermal gene, rhomboid (rho). The *rho* gene has an expression pattern similar to that of the ind gene and its enhancer is well studied; in fact, Sayal et al. [18] experimentally characterized the expression pattern driven by this enhancer as well as 37 synthetic variants thereof. Since the ind enhancer (subject of our modeling above) and the

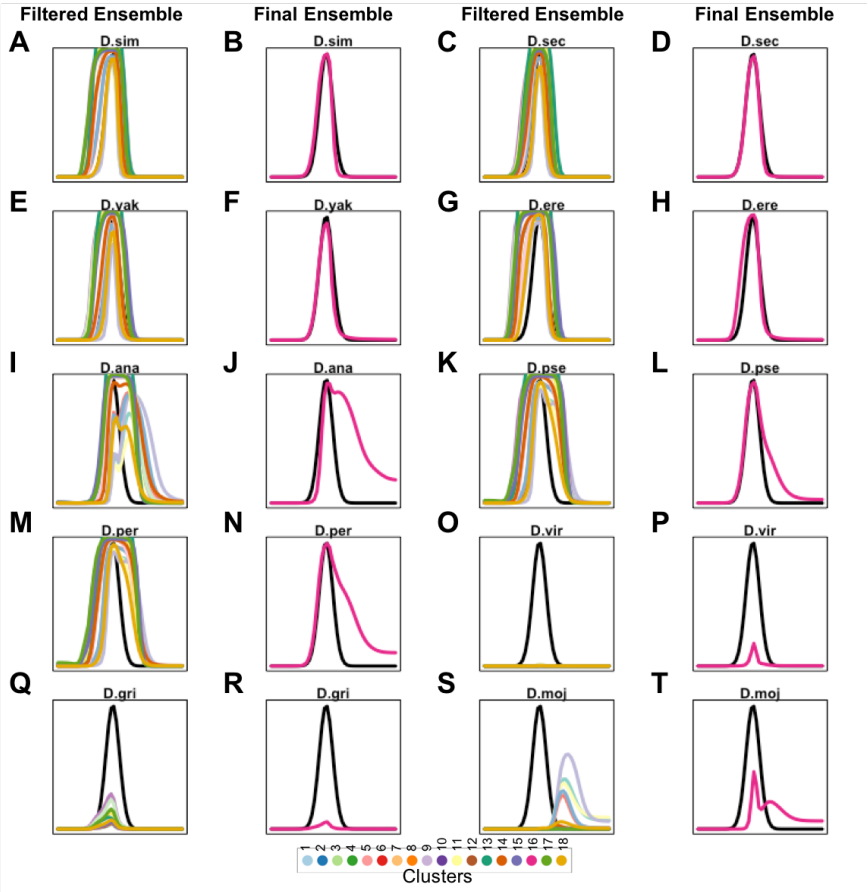

Figure 3.6: Cross validation of models on orthologs of ind enhancer. (A-T) For each orthologous enhancer sequence, we predicted the expression profile using models trained on D. melanogaster. Predicted expression profiles from models in each cluster of the ensemble were averaged and are shown by a line in a cluster-specific cluster. Averaged predictions from the cluster known as the final ensemble are separately shown in panels in the second and fourth columns (magenta). Orthologs were taken from D. simulans (A,B), D. sechellia (C,D), D. yakuba (E,F), D. erecta (G,H), D. ananassae (I,J), D. pseudoobscura (K,L), D. persimilis (M,N), D. virilis (O,P), D. grimshawi (Q,R), D. mojavensis (S,T). We observe that for orthologs in D. mojavensis and D. virilis, two of the most distantly related enhancers, the final ensemble (magenta) predicts expression in the correct position along the D/V axis (same as in D. melanogaster) while other groups of models misplace the peak of expression (for D. mojavensis) or predict no expression (D. virilis). (U) Mean RMSE ("root mean squared error") between wild-type ind expression and the expression profile predicted for the orthologous enhancer from each species, for each cluster of models. We noted that the mean RMSE for the final ensemble (magenta dots) is at the lower end of the distribution for each species.

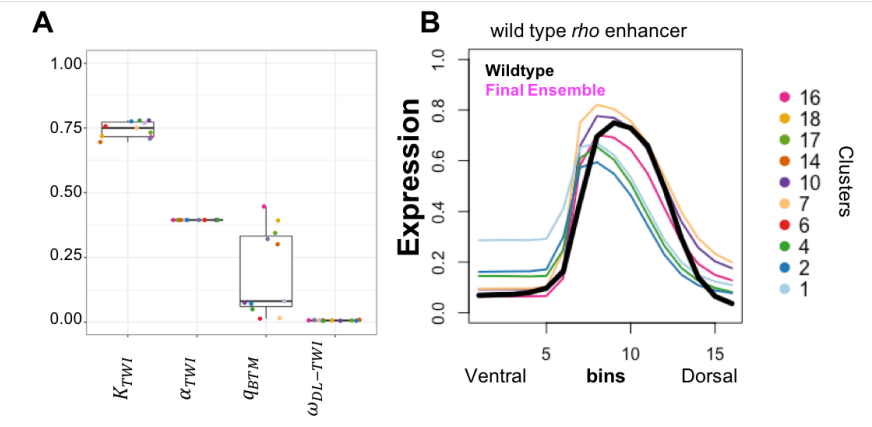

Figure 3.7: Cross validation on *rho* enhancers. We used models of ind data as starting points to train models for the wild-type *rho* enhancer. Four of the model parameters represent TFs relevant to both ind and *rho* and were kept fixed at values trained on ind data; four additional parameters were defined exclusively for modeling *rho* and needed to be trained. Models in eight of the 18 ensembles in the filtered ensemble for ind data were incompatible with *rho* data, and all models in the newly trained ensemble for *rho* data corresponded to (shared parameter values with) the remaining 10 clusters in the ind ensemble, including the "final ensemble". (A) Distributions of values learned for the four new parameters are shown as box plots, where each dot represents the average parameter value of models in a specific cluster. (The points are colored by cluster.) (B) Expression profile of wild-type *rho* enhancer as predicted by the models. The average predicted profile from models in each cluster is shown individually, in cluster-specific colors, with the magenta line representing prediction by the "final ensemble" (cluster 16).

rho enhancer have similar expression patterns (outputs) and share regulators (inputs), we sought to cross-validate our models, trained with ind data, on the *rho* enhancer and its variants [18].

The *rho* enhancer is known to be controlled by two activators – DL and TWI – and one repressor, SNA. While DL and SNA were among the TFs included in the models of ind above, TWI was not, and as a result the trained models are not capable of predicting *rho* expression. To address this, we performed partial optimization of parameters on the *rho* data set (37 synthetic constructs) from Sayal et al [18]. In particular, we considered each model trained on ind data (previous sections), utilized the trained values of four of its parameters that are shared between ind and *rho* models without further modification, but trained four additional parameters unique to *rho* on the *rho* data set (see Methods). As a result, each model in the filtered ensemble from above gives rise to a model for the *rho* data, with four unchanged parameters and four newly trained parameters (Figure 3.7 A). The accuracy of the resulting model on each of the *rho* enhancers (wild type and its variants)

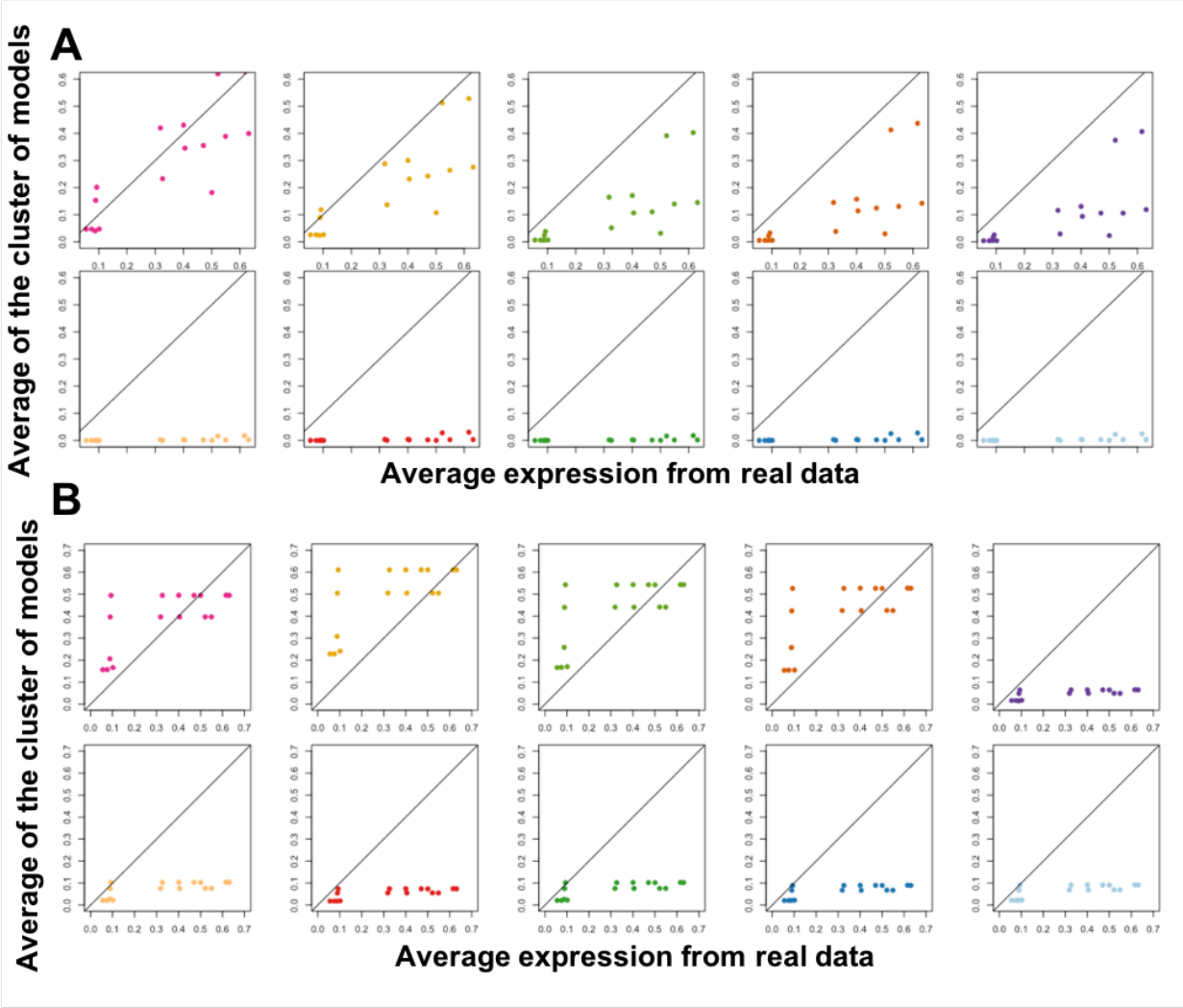

Figure 3.8: (A) Average expression in the bins near the peak of the *rho* expression profile along the D/V axis (bins 8-12). The x-axis is the average expression from experimental data and the y-axis is the model-predicted average expression in this domain. Each dot is a construct (wild-type *rho* enhancer or variant thereof) in the dataset. Each panel is the from a separate cluster of models, points being in a cluster-specific color (same as panel B). The top left panel represents the final ensemble. (B) Similar to C, but the average expression represents the ventral-most five bins (bins 1 to 5). The y-axis is the model-predicted expression (average in this domain) and the x-axis represents the experimental data.

was then assessed using SSE score. We noted that not all models in the filtered ensemble led to models capable of explaining the *rho* data set; rather, only models belonging to 10 of the 18 groups of models in the ensemble could, upon training of the additional parameters, provide fits better than a modest threshold of SSE = 0.1. The prediction of each of these 10 groups of models for the wild-type *rho* enhancer is shown in Figure 3.7B.

We next examined predictions of the above models on 26 enhancers that differed from the wild-type *rho* enhancer in that the peak expression driven by these variant enhancers is significantly lower than that of the wild-type enhancer (Supplementary Figure A.11). We assessed how well the models in each group from the filtered ensemble capture this phenomenon: for each group of models we computed the average predicted expression in the peak expression region and compared it to the true (experimentally measured) expression in that region (Figure 3.8 A). It was visually clear that models from the final ensemble (top left panel in Figure 3.8 A) captured the reduced peak expression levels of these 26 variant enhancers better than all other groups of models of the filtered ensemble. We similarly examined predictions on the 9 enhancers that differ from the *rho* enhancer in a clear derepression in the ventral-most region of the embryo (Supplementary Figure A.12). For these enhancers, we computed the predicted expression, from each group of models, in this spatial region and compared it to the experimentally observed expression in the region (Figure 3.8 B). Visual inspection reveals that the models originating in the final ensemble capture the phenomenon of ventral derepression better than six of the other groups of models and at least as well as the remaining three groups in the filtered ensemble. In summary, the final ensemble identified above based on our experimental assessment of mutation effects proved to be far more accurate than competing groups of models, in terms of its ability to generalize to a new data set comprising a related but distinct enhancer (the *rho* enhancer) and its variants.

3.8 DISCUSSION

A major open problem today is how DNA sequence variations, e.g., single-nucleotide polymorphisms (SNPs), lead to phenotypic differences among individuals. A popular approach is to find polymorphisms that are statistically correlated with the phenotype, as in genome-wide association studies (GWAS) [5], family-based association tests [88], and expression quantitative trait loci (eQTLs) [89, 90] for phenotype-related genes. However, statistically identified variations may not be functionally related to phenotypes [91], due to a variety of factors including linkage disequilibrium (LD) and redundancy of genetic systems. This problem is particularly pronounced in the case of non-coding variations, which form the majority

of GWAS findings [4] and function by influencing gene regulation. Accurate contextual or mechanistic information about non-coding variations can help us pinpoint those that are causally related to the phenotype [80, 6]. The work presented here is a step in this direction, and provides an example of how detailed mechanistic models of the sequence-to-expression relationship encoded by an enhancer may help us predict the effects of non-coding variations. This in turn can lead to better prioritization of phenotype-related variants and also provide mechanistic explanation of their effects.

In recent years, various machine learning-based methods such as gkm-SVM [28] and DeepSEA [27] have been proposed for modeling the sequence-function relationship encoded throughout the non-coding genome. These have been successful in predicting the impact of variants on epigenomic states such as DNA accessibility and TF-DNA binding [28, 27], although some reports indicate there is significant room for improvement in their accuracy [92] There is also evidence that these machine learning methods can help identify eQTLs and disease-related variations. The thermodynamics-based modeling approach used in our work offers a complementary approach to variant interpretation, and while it is far less scalable than the ML-based methods it is more mechanistically grounded and potentially more precise. It is also possible that similar models, once trained on high throughput data such as those from massively parallel reporter assays [93], will provide mechanistic predictions about non-coding variations on a larger scale than the current work. In a simple illustration of how mechanistic models can be useful at scale, Xie et al. [94] showed that motif-based biophysics-inspired models of TF-DNA binding predict SNPs that likely impact binding strength and lead to inter-individual variation in chemosensitivity.

The use of mechanistic quantitative models to examine polymorphisms, though uncommon today, is not entirely new. Gursky et al. [21] analyzed polymorphisms in Drosophila strains (the same collection as in our study) using the same sequence-to-expression model. They reported several valuable insights, including additive effect of multiple polymorphisms in individual genotypes and evidence of selective pressure at the level of combinations of SNPs. Our analysis has conceptual and methodological similarities to Gursky et al., with a few key differences. First, our approach recognizes that uncertainties in the model (values of parameters) can lead to ambiguities in polymorphism analysis, and our predictions are accompanied by estimates of the resulting uncertainty. More importantly, while Gursky et al. focused primarily using the sequence-to-expression model to reveal insights about a collection of polymorphisms, we focus more on functional analysis and use experimental assays of variant effects to refine the models, making them more precise and more ready for future applications. Our work also underscores the value of model-based design of biological experiments. Two of the three enhancer variants that we tested were chosen because models

based on prior data were ambiguous in their predictions regarding those variants. After we performed those experiments, the results led to a significant narrowing of the feasible models and this smaller feasible group of models was then shown to be more consistent with held-out data sets (based on orthologs of the ind enhancer as well as several synthetic variants of the *rho* enhancer) than the original broader ensemble of models. We hope that such iterative applications of modeling and experimental testing, with models furnishing candidates for experimentation and experimental results refining the models, will be more frequently adopted in future investigations.

CHAPTER 4: TISSUE SPECIFIC ANALYSIS OF NON-CODING VARIATIONS

4.1 STUDYING GENETIC VARIATIONS IN THE NON-CODING DNA

Analyzing genetic variants in the genome is appealing as it can unveil not only the genetic basis of disease, but in studying interpersonal genetic variation, imparts knowledge as to how SNPs and other genetic variations exacerbate the fractional risk of developing that disease. For instance, the neuro-degenerative disease dementia is not wholly inherited, but is partly influenced by variation associated with the particular SNPs a person has inherited [95]. Identifying genetic variants with functional impacts is thus essential to advancing precision medicine tailored to individuals. The most challenging task in studying genetic variation lies in assessing their functional impact on direct molecular interactions and the manner in which these effects cascade to influence systemic cellular and organismal processes.[96]. One particular cellular process that genetic variants have clear and significant influence over is gene regulation, the phenomena in which the protein products of one gene bind to preferred DNA sequences proximal to a target gene in order to attenuate the kinetics of the transcriptional expression of the target). The composition and distribution of these preferred DNA sequences (binding sites) of a given transcriptional regulatory factor (TF) can be disrupted by alterations in the TF binding sites (TFBS). Understanding the mechanisms and underlying logic governing gene regulation thus, is critical to interpret the exact consequence of genetic variants on such an instrumental cellular process.

Unlike the process of DNA replication or transcription itself, the biology of gene regulation is far from completely understood; further compounding the problem is that the one known characteristic of this phenomena, that regulators bind in noncoding regions of DNA, is accompanied by the fact that the majority of inherited SNPs lie in the non-coding regions which harbor the regulatory elements. Though such SNPs can disrupt the effectiveness of TFBS, the overlapping of SNPs within TFBS is not sufficient to conclude a regulatory effect, making TFBS as a functional filter much more difficult to wield than say non-synonymous coding SNPs [97]. The consequences of sequence variation depend on a litany of circumstances, including the tissue in which the gene is functional, the regulatory network or functional pathways in which the gene is involved, the organism developmental stage, the exact base pair and position in the TFBS the SNP interrupts, the behavior of other regulatory factors, the precise logic of the regulation between a TF and target gene, among countless other complicating variables. [7, 17, 16]. Even concentrating on this interpreting genetic variation in the context of this particular biological process incurs significant

complication. Nevertheless, researchers have developed statistical and mathematical models to, piece by piece, uncover the causal links between the DNA variations and gene expression levels in individuals [96, 6, 17].

Genome-wide association studies (GWAS) have identified thousands of single-nucleotide variants associated with complex traits or diseases [98, 99]; however, the functional impact of these variants and their connections to disease etiology remain elusive. GWAS, and association methods in general that try to implicate genetic variants in isolation of one another, run into problems with linkage disequilibrium (LD), a phenomena in which proximal genetic variants are likely to covary; because SNPs close to one another correlate, correlation based methods like GWAS that implicate one SNP in an LD block tend to implicate them all, which means the strength of the statistical association is insufficient to infer a biological mechanism without additional context.

The genetic variants that underlie the phenotype are more likely to be associated with the gene expression (i.e. expression quantitative trait loci (eQTL)) [100, 101, 102]. One major issue in these studies is that some may not be causal although studies find them significant. Modeling the entire system of biological entities and their relations is computationally prohibitive and data intensive and so elucidating the set of transcriptional regulatory dependencies among genes presents a feasible and yet powerful alternative to identify the cellular processes pertinent to disease biology. One of the simplest and most common methods for constructing GRNs relies on a general technique called co-expression analysis [103, 104, 105], which identifies clusters of genes that covary and, from these clusters, uses regression analyses to infer where clusters of genes are identified on the basis of similarities in expression profiles across samples; from these clusters, regression analyses can help infer network topology, where the nodes of the network correspond to genes and directed edges, from say gene A to gene B, imply gene A regulates the expression of gene B. Such co-expression procedures suffer from myriad problems when not supplemented with additional information. For instance, co-expression analyses omit much of the underlying biology of transcriptional regulation, assuming that correlated gene expression is proof of a regulatory relationship. However, this need not be true as other factors can account for co-expression, such as cellular perturbations and cell cycle. In cases where these are well controlled, straight forward co-expression methods produce convolved network structures that disregard the sequence of regulation; for example, if gene A regulates B which regulates C, convolved networks suggest that gene A directly regulates gene C. Network deconvolution methods are then required to infer the real GRN network structure from its noisy reconstruction, often through latent variable analyses [106]. Such methods though have their own limitations and recent research [107].

A staple of regulatory biology is that regulators bind in proximity to the genes they regulate and in doing so, change the kinetics of transcription to influence the target gene's expression. By not taking into consideration the distribution and strength of TF binding sites near target genes, co-expression derived GRNs fall prey to false positives (i.e. illegitimate TF $\rightarrow$ target gene links) as they may be missing the needed biology for the TF to exert its influence. This is but one of many ways in which coexpression itself is insufficient for accurate GRN reconstruction. We employed a probabilistic graphical model to construct a gene regulatory network and find mutations that mediate the effect of gene expression variation across individuals. We used GTEx data [108] to build a tissue specific regulatory network and predict the mutations that causally influence the gene expression levels through changing binding sites of relevant TFs. To our knowledge, this represents the first such approach to combine the knowledge of TF binding and the single nucleotide variations associated with the gene expression changes (eQTL SNPs) as regulatory evidence to de novo construct transcriptional regulatory networks specific to the tissue.

4.2 ANALYSIS OF SNPS LEVERAGING GENE REGULATORY EVIDENCE

Gene regulatory network can control which genes are expressed and the extent of gene expression levels. There are many factors that can decide the transcription of the genes and the protein products, such as transcription factors and other protein products [109, 110]. Understanding the role of gene regulatory network is the fundamental step in shaping a mechanistic view and interpreting the effect of genetic variations. Variation in gene expression levels is co-related with changes in the phenotype and the disease state. Transcription factors (TFs) regulate gene expression levels and perturbations in the binding sites for the TFs that regulate the gene may cause a change in the gene expression. Minor changes such as a single nucleotide change in the DNA can potentially affect the gene regulatory mechanisms.

Assumptions of the model:

- The TF regulates certain genes to influence the phenotype
- TF does not change its role as an activator or a repressor in a given tissue
- Perturbing the relationship (e.g., by changing the strength of the binding site) should impact expression of the genes targeted by the TF

The eQTL data provides information on the variants that associate with the expression changes (figure 4.1 A). We use the GRN to assess how regulatory processes might be per-

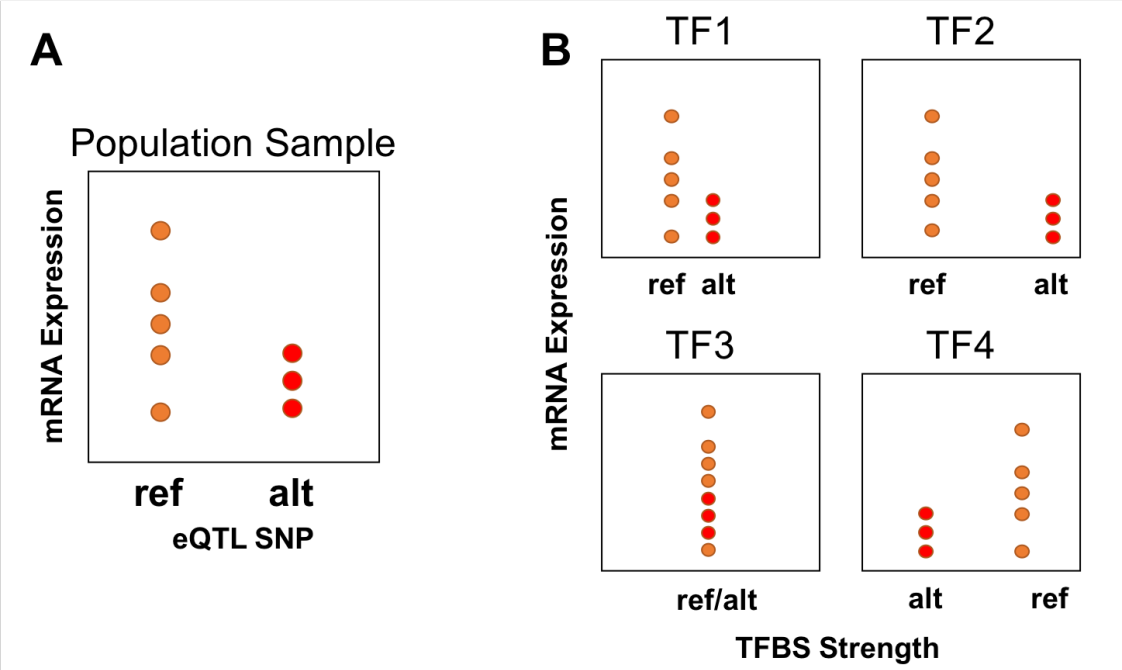

Figure 4.1: **Relationship between gene expression and genotype** These panels are schematic diagrams of gene expression variations across different individuals. Assuming we have mRNA levels of a given gene for many individuals and a SNP position where there is a difference between gene expression levels of individuals carrying the SNP. A) The eQTL analysis finds SNPs with significant associations. The y-axis is the gene expression level and the x-axis is the genotype. Each point on the scatter plot represents gene expression level of one individual. B) The x-axis represents the change in binding strength of the transcription factors at the location of the SNP. TF1 has smaller change in the TFBS than the TF2 but they play the same role in regulating the gene. TF1 and TF2 act as a repressor since increasing the binding strength of the TF decreases the gene expression levels. TF3 does not have an effect and TF4 plays the opposite role (activator) in the gene regulation.

turbed and determine the role of SNPs in this process. From the GRN, we can determine if the TF targets the gene and the GRN can be constructed using the data from the strength of the binding sites of the TF and the significance of the eQTL p-values. If the TF that targets the gene is an activator, increasing the transcription factor binding strength (TFBS) increases the gene expression levels and decreasing the TFBS decreases the expression levels (figure 4.1 B).

In this work, we provide a simple statistical framework for associating transcriptional mechanisms with gene expression variations across different individuals. The algorithm proposes that certain SNPs mediate the effect of transcription factors on gene expression variations. More specifically, we assume the effect is mediated through the SNPs that are within 100 Kbp of transcription start sites of gene's and their presence correlates with the

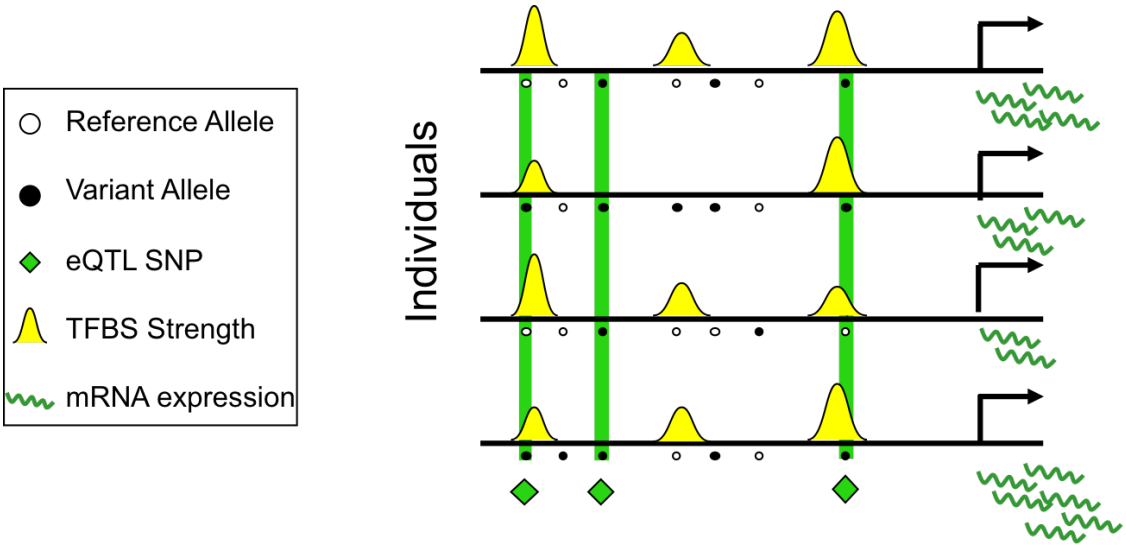

Figure 4.2: **Diagram of data** Single Nucleotide Polymorphisms (SNPs) are depicted by white or black dots depending on the allele. The gene expression levels are shown as the green squiggles. For a given TF, the strength of the binding is measure by the yellow peaks. The green diamonds depict the positions of the eQTLs and the green horizontal lines are the functional positions of the genome.

expression variations of the genes (eQTL SNPs) and change the binding sites of associated TFs. While these signals can be noisy in gene levels, aggregating them through different targets genome wide allows the true signal to be detected. In this manner, we can interpret SNPs in a cis-regulatory context and link them to genes in a biologically meaningful manner. The following is the model we propose for this study. We use motifs to scan the DNA positions with the genetic mutation reported in the original data. We use a score that reflects the amount of change in the TFBS due to the genetic mutation. We represented a schematic of the data that the model aggregates in figure 4.2. The change in TFBS is depicted in yellow, the eQTL SNPs are shown with green diamonds. The functional locations are highlighted by vertical green lines.

4.2.1 Data Collection

In this study, we combined several pieces of information from different studies to define the genotype- phenotype relationship: (1) The List of genes for which significant eQTLs are available from GTEx data; (2) CIS- BP dataset of DNA binding specificity for about 700 different TFs [111]; and The goal of the Genotype-tissue expression (GTEx) project is to estimate how the genetic variants change the gene expression in different tissues [112, 113]. The data is collected from 485 human donors across 48 tissues and the expression is measured

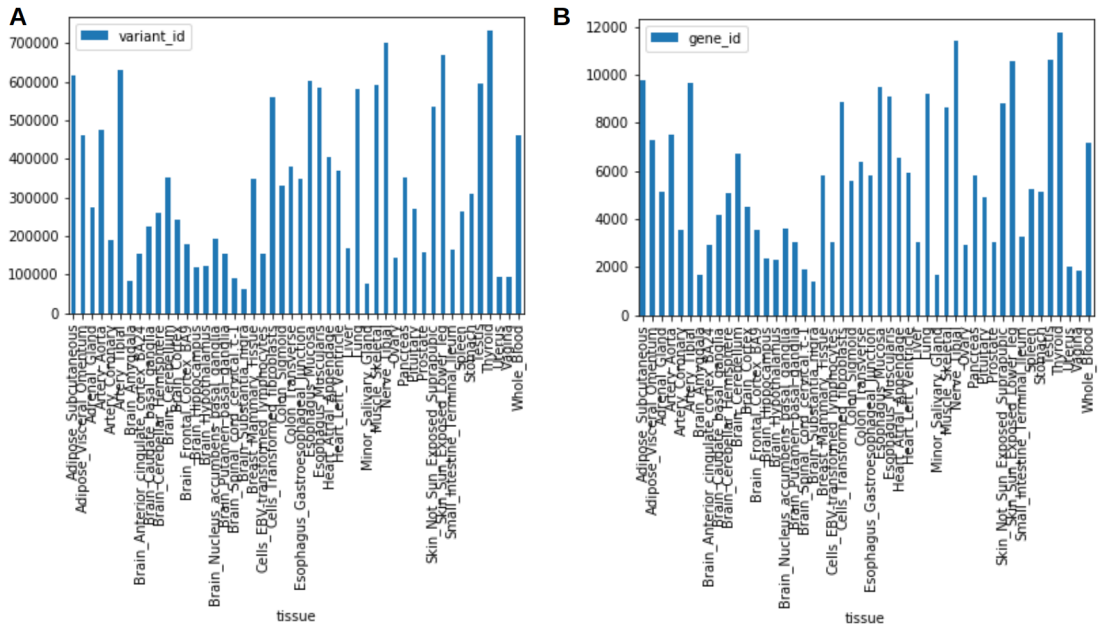

Figure 4.3: **GTEx eQTL variants and gene for 48 different tissues** A) Bar plot of the number of unique eQTL SNPs for each tissue. B)Bbar plot of number of unique eQTL genes for each tissue

for every transcript in the samples (˜55,000 transcriptome)(Figure 4.3). We take the SNP positions for each tissue and we scan the location of the genome with the reference allele and the alternative allele [114]. Each motif is a position weight matrix (PWM) that specifies the preference of TF to bind to the DNA. We use 1841 experimentally confirmed PWMs for 700 different transcription factors from the CIS-BP dataset [105, 104].

4.3 A PROBABILISTIC MODEL TO INTEGRATE REGULATORY EVIDENCE WITH EXPRESSION QUANTITATIVE TRAIT LOCUS

The data from the Genotype-Tissue Expression (GTEx) project provides a large set of SNP-gene associations identified via expression quantitative trait loci (eQTL) analysis - a procedure that fractionally attributes the genetic variance of a gene's expression to some loci [115], often a single SNP (called an eQTL SNP if its eQTL analysis is sufficiently statistically significant). The multicollinearity of SNPs proximal to a gene make it impossible to disambiguate which SNP may be functionally responsible for the observed variance in gene expression, and thus insufficient as a basis to infer causality.

Our goal was to address this issue by developing a principled method that could identify the functional SNPs for a gene by examining, simultaneously, all eQTL SNPs for a gene and

their strengths of eQTL association. To aid in this pursuit, we supplemented the eQTL data with ancillary data on the regulatory impact of each SNP on a TF's binding to DNA (via the TF's motif). In doing so, we re-frame the aforementioned eQTL analysis around the biology of transcription factors and transcriptional regulation. Consequently, the functional effects we hope to elucidate observed correlations between genetic variation and gene expression variation are regulatory in nature. Although we cannot promise to uniquely attribute a single SNP to a gene's expression, by conducting the analysis in a regulatory context, we hope to significantly reduce the number of viable candidates.

After careful consideration, we elected to use probabilistic graphical models (PGMs) to address this challenge - as such models provide a language to describe latent properties of the data, such as the regulatory state of each SNP and TF, and tools to infer such states on the basis of empirical evidence. In designing this PGM, shown in figure 4.4, we make the following assumptions. The PGM we propose here considers a single TF model at a time. For such a model, it is assumed that the TF regulates each gene independently, but that the regulatory effect of a TF on a gene (activation/repression) is binary (true/false), latent, and unknown at the time of inference. The model considers two branching paths depending on if the TF regulates the gene. In cases where the TF regulates the gene, we expect the p-value of the eQTL and p-value of the impact of the SNP on motif match strength to be small for certain SNPs nearby the gene. However, in cases where the TF does not regulate the gene, we expect our evidences to be independent of one another.

4.3.1 Overview

We evaluate the probabilistic model, described in figure 4.4, separately for each TF. The model integrates different sources of regulatory evidence of the TF targeting the gene and the SNP simultaneously disturbs the TF binding site and the gene's expression. We rely on regularly evidences of the TF regulating the gene, each weighted in accordance with its contribution to the model, which combine in a probabilistic framework to determine the probability the gene is a target of the TF and certain SNPs of the gene mediate the effect of the TF on gene expression. In this analysis, we use two regulatory evidence: Motif scores of the TF and from the GTEx data cis-eQTLs.

For each gene g in a tissue t, we get TFs indexed by T, such that there is an edge in the GRN between TF_T and gene g . For each pair of (TF_T, g), the probabilistic graphical model (PGM) determines if there is an edge between TF_T and gene g and whether SNP s nearby the gene g mediates the effect of T on the expression of g. Two hidden binary variables Z_g and S_{ig} capture these relationships where $Z_g = 1$ if T targets g and $S_{ig} = 1$

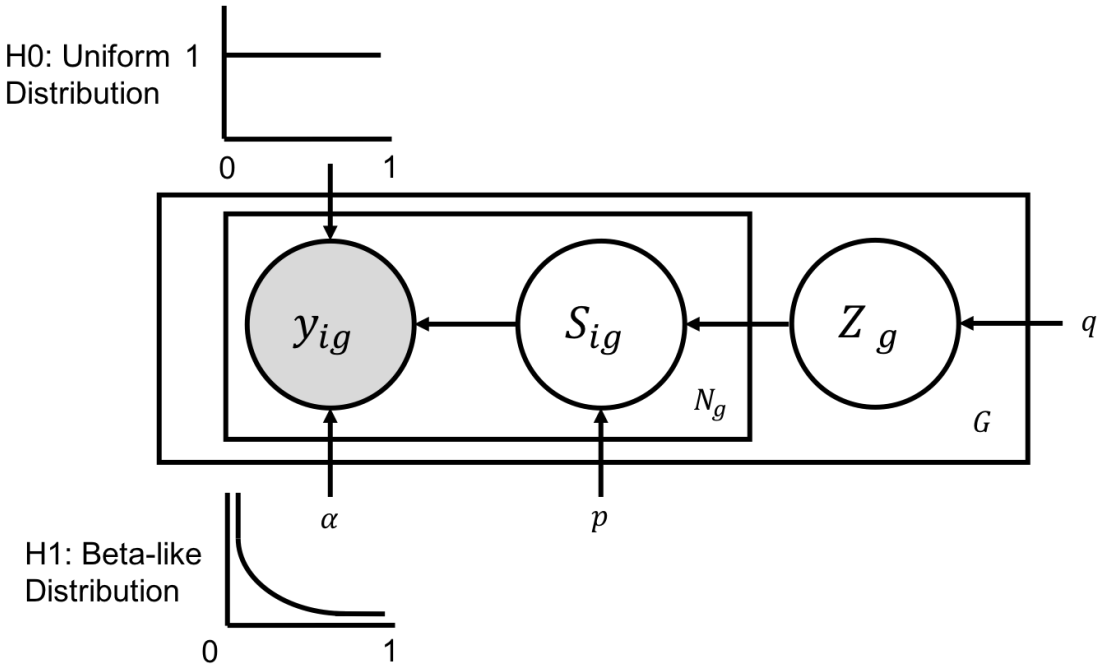

Figure 4.4: **Plate diagram for the PGM model.** In this diagram, latent variable Z_g represents whether the TF targets a gene g and its enclosing rectangle denotes G such genes. The latent variable S_{ig} shows whether the SNP i associated with the gene g, mediates the effect of the TF on the gene's expression; and the rectangle represents N_g number of SNPs for each gene g. If $Z_g = 1$ and $S_{ig} = 1$, we expect an enrichment for significant eQTL p-values and significant changes in the LLR score of the motif. The joint distribution for p_{motif}, p_{eQTL} is modeled by a beta distribution with parameter $\alpha < 1, \beta = 1$; otherwise,the model expects uniformly distributed p-values in [0,1] (Null hypothesis H0).

if the i'th SNP associated with the gene g mediates the effect of T on g. There are four parameters σ, α, p and q (see below for explanations of these parameters). Training data is $y_{ig} = \frac{p_{motif,ig}+p_{eQTL,igt}}{2}$ where $p_{motif,ig}$ denotes the p-value of the motif score for the TF in a window around the i SNP position of gene g and $p_{eQTL,igt}$ is the eQTL p-value nominal from the GTEx data for the tissue t at the i'th SNP of the gene g. There are 1.6 million motif p-values and the number of eQTL p-values varies accross different tissues.

4.3.2 Detailed Description

N := Number of genes.
N_g := Number of SNPs within TFBS for gene.
$p_{eQTL,igt}$:= The eQTL p-value for the i-th SNP and gene g in tissue t
σ := +1 if activator, -1 if repressor (set beforehand)

$p_{motif,ig}$:= The 1 – CDF of motif score of the i-th SNP of the gene g (reference vs alternative allele).

$y_{ig} = \frac{p_{motif,ig} + p_{eQTL,igt}}{2}$

Given all y_{ig} in tissue t as Y, we estimate the likelihood of Y as the following:

$$P(Y|p,q,\alpha) = \prod_{g=1}^{G} \sum_{j=0}^{1} P(Z_g = j|q,\alpha) \prod_{i=1}^{N_g} \sum_{k=0}^{1} P(y_{ig}, S_{ig} = k|Z_g = j,q,\alpha) \times \mathbb{1}(Z_g = j, S_{ig} = k) \tag{4.1}$$

The $\mathbb{1}(x)$ is the identity function where it evaluates to 1 if the statement x is true and otherwise, it is 0.

Z_g := A binary latent variable representing whether or not the correlation p-value of gene g's expression with the phenotype of interest is drawn from a uniform or beta distribution.

α := Parameter determining the shape of a beta distribution. The β parameter is set to 1 and α is capped in the range [0,1]. This parameter is estimated by the model.

p_g := A continuous observed variable in the range [0,1] representing the correlation p-value of gene g's expression with the phenotype of interest obtained from the GTEx data.

$$\begin{aligned} P(Y|p,q,\alpha) = \prod_{g=1}^{G} \Bigg[& \\ P(Z_g = 0|q) \prod_{i=1}^{N_g} \times \mathbb{1}(Z_g = 0, S_{ig}) \Big\{ & P(y_{ig}|S_{ig} = 0, Z_g = 0, q, \alpha) \times P(S_{ig} = 0|Z_g = 0, q, \alpha) \\ & + P(y_{ig}|S_{ig} = 1, Z_g = 0, q, \alpha) \times P(S_{ig} = 1|Z_g = 0, q, \alpha) \Big\} \\ P(Z_g = 1|q) \prod_{i=1}^{N_g} \times \mathbb{1}(Z_g = 1, S_{ig}) \Big\{ & P(y_{ig}|S_{ig} = 0, Z_g = 1, q, \alpha) \times P(S_{ig} = 0|Z_g = 1, q, \alpha) \\ & + P(y_{ig}|S_{ig} = 1, Z_g = 1, q, \alpha) \times P(S_{ig} = 1|Z_g = 1, q, \alpha) \Big\} \\ \Bigg] & \end{aligned} \tag{4.2}$$

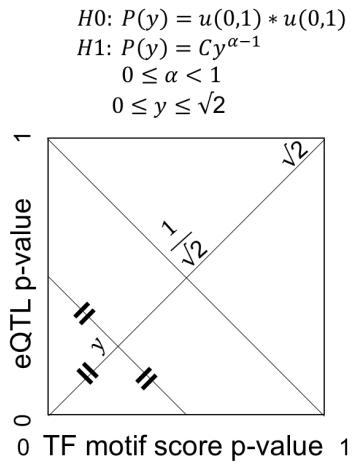

Figure 4.5: Diagram of integration

4.3.3 Likelihood

The probability that $Z_g = 1$ is a Bernoulli function:

$$P(Z_g = 1|q) = q \tag{4.3}$$

The probability of $S_{ig} = 1$ given that the TF regulates the gene is:

$$P(S_{ig} = 1|Z_g = 1) = p \tag{4.4}$$

The probability of $S_{ig} = 1$ given that the TF does not regulate the gene is:

$$P(S_{ig} = 1|Z_g = 0) = 0 \tag{4.5}$$

If $S_{ig} = 1$, then $P(y_{ig}) = \frac{\alpha+1}{(1-2^{-\alpha})} \times \alpha {y_{ig}}^{\alpha-1}$

$$\boxed{pdf(y_{ig}) = \begin{cases} C \times \alpha {y_{ig}}^{\alpha-1} \times y_{ig} & \text{if } 0 \leq y_{ig} \leq \frac{1}{2} \\ C \times \alpha {y_{ig}}^{\alpha-1} \times (1 - y_{ig}) & \frac{1}{2} \leq y_{ig} \leq 1 \end{cases}} \tag{4.6}$$

where $C = \frac{\alpha+1}{(1-2^{-\alpha})}$

Proof:

$$1 = C \times \left[\int_0^{\frac{1}{2}} \alpha y^{\alpha-1} y dy + \int_{\frac{1}{2}}^1 \alpha y^{\alpha-1}(1-y)dy \right] \tag{4.7}$$

$$= C \times \left[\int_0^{\frac{1}{2}} \alpha y^{\alpha} y dy + \int_{\frac{1}{2}}^1 \alpha y^{\alpha-1} dy - \int_{\frac{1}{2}}^1 \alpha y^{\alpha} dy \right] \tag{4.8}$$

$$= C \times \alpha \left[\frac{1}{\alpha+1} \times (\frac{1}{2})^{\alpha+1} - 0 + \frac{1}{\alpha} - \frac{1}{\alpha} \times (\frac{1}{2})^{\alpha} - \frac{1}{\alpha+1} + \frac{1}{\alpha+1} \times (\frac{1}{2})^{\alpha+1} \right] \tag{4.9}$$

$$= \frac{C}{\alpha+1} \times (1 - 2^{-\alpha}) \tag{4.10}$$

If $S_{ig} = 0$, then

$$P_{null}(y_{ig}) = P(y_{ig}|S_{ig} = 0) \approx uniform \tag{4.11}$$

$$P(Y|p,q,\alpha) = \prod_{g=1}^{G} \left[\mathbb{1}(Z_g = 0) \times (1-q) \times \prod_{i=1}^{N_g} \left\{ P_{null}(y_{ig}) \times 1 + P_{alt}(y_{ig}) \times 0 \right\} + \right.$$
$$\left. \mathbb{1}(Z_g = 1) \times q \times \prod_{i=1}^{N_g} \left\{ P_{null}(y_{ig}) \times \mathbb{1}(S_{ig=0})(1-p) + \mathbb{1}(S_{ig} = 1) \times p \times P_{alt}(y_{ig}) \right\} \right] \tag{4.12}$$

$$\boxed{P(Y|p,q,\alpha) = \prod_{g=1}^{G} \left[(1-q) \prod_{i=1}^{N_g} P_{null}(y_{ig}) + q \prod_{i=1}^{N_g} \left\{ P_{null}(y_{ig}) \times (1-p) + P_{alt}(y_{ig}|\alpha) \times p \right\} \right]} \tag{4.13}$$

$$P(Y|p,q,\alpha) = \prod_{g=1}^{G} \left[(1-q) \prod_{i=1}^{N_g} P_{null}(y_{ig}) + q \prod_{i=1}^{N_g} \left\{ P_{null}(y_{ig}) \left((1-p) + \frac{P_{alt}(y_{ig})}{P_{null}(y_{ig})} \times p \right) \right\} \right] \tag{4.14}$$

$$P(Y|p,q,\alpha) = \prod_{g=1}^{G} \left[(1-q) \prod_{i=1}^{N_g} P_{null}(y_{ig}) + q \left\{ \prod_{i=1}^{N_g} \left(1 - p + \frac{P_{alt}(y_{ig})}{P_{null}(y_{ig})} \times p \right) \times \prod_{i=1}^{N_g} P_{null}(y_{ig}) \right\} \right] \tag{4.15}$$

$$P(Y|p,q,\alpha) = \prod_{g=1}^{G}\left[(1-q) + q \times \prod_{i=1}^{N_g}\left(1 - p + \frac{p \times P_{alt}(y_{ig})}{P_{null}(y_{ig})}\right)\right] \times \prod_{g=1}^{G}\prod_{i=1}^{N_g} P_{null}(y_{ig}) \quad (4.16)$$

$$\log P(Y|p,q,\alpha) = \sum_{g=1}^{G} \log\left[(1-q) + q \times \prod_{i=1}^{N_g}\left(1 - p + \frac{p \times P_{alt}(y_{ig})}{P_{null}(y_{ig})}\right)\right] + \sum_{g=1}^{G}\sum_{i=1}^{N_g} \log P_{null}(y_{ig}) \quad (4.17)$$

$$\boxed{\log P(Y|p,q,\alpha) = \sum_{g=1}^{G} \log\left[(1-q) + q \times e^{\sum_{i=1}^{N_g} \log\left(1-p+\frac{p\times P_{alt}(y_{ig})}{P_{null}(y_{ig})}\right)}\right] + \sum_{g=1}^{G}\sum_{i=1}^{N_g} \log P_{null}(y_{ig})} \quad (4.18)$$

Let $A = \log(1-q)$ and $B = \log(q) + \sum_{i=1}^{N_g} \log\left(1 - p + \frac{p\times P_{alt}(y_{ig})}{P_{null}(y_{ig})}\right)$ then:

$$\log P(Y|p,q,\alpha) = \sum_{g=1}^{G} logSumExp(A,B) + \sum_{g=1}^{G}\sum_{i=1}^{N_g} \log P_{null}(y_{ig}) \quad (4.19)$$

Note that we can write $x = e^{\log(x)}$ and $\log(\prod x_i) = \Sigma_i \log(x_i)$ The function $logSumExp(x,y) = \log(e^x + e^y)$.

We can either optimize this function directly or use Expectation Maximization. In this formulation, we choose EM. Then, we compare the log likelihood ratios between this model and the null model to get a ranking of different transcription factors in each tissue.

4.3.4 Expectation Maximization

Let $\theta = (q, \alpha, p)$ then

$$Q(\theta^t|\theta^{t-1}) = \sum_{k=0}^{1}\sum_{j=0}^{1}\sum_{g=1}^{G}\sum_{i=1}^{N_g} P(Z_g = k, S_{ig} = j|y_{ig}, \theta^{t-1}) \log P(y_{ig}, Z_g = k, S_{ig} = j|\theta^t) \quad (4.20)$$

$$Q(\theta^t|\theta^{t-1}) = \sum_{g=1}^{G}\sum_{i=1}^{N_g} P(Z_g = 0, S_{ig} = 0|y_{ig}, \theta^{t-1}) \log P(y_{ig}, Z_g = 0, S_{ig} = 0|\theta^t) \tag{4.21}$$
$$+ P(Z_g = 0, S_{ig} = 1|y_{ig}, \theta^{t-1}) \log P(y_{ig}, Z_g = 0, S_{ig} = 1|\theta^t)$$
$$+ P(Z_g = 1, S_{ig} = 0|y_{ig}, \theta^{t-1}) \log P(y_{ig}, Z_g = 1, S_{ig} = 0|\theta^t)$$
$$+ P(Z_g = 1, S_{ig} = 1|y_{ig}, \theta^{t-1}) \log P(y_{ig}, Z_g = 1, S_{ig} = 1|\theta^t)$$

$$Q(\theta^t|\theta^{t-1}) = \sum_{g=1}^{G}\sum_{i=1}^{N_g} w_{00,ig}^{t-1} \log\left(P(y_{ig}|S_{ig} = 0, \theta^t) \times P(S_{ig} = 0|Z_g = 0, \theta^t) \times P(Z_g = 0|\theta^t)\right) \tag{4.22}$$
$$+ w_{01,ig}^{t-1} \log\left(P(y_{ig}|S_{ig} = 1, \theta^t) \times P(S_{ig} = 1|Z_g = 0, \theta^t) \times P(Z_g = 0|\theta^t)\right)$$
$$+ w_{10,ig}^{t-1} \log\left(P(y_{ig}|S_{ig} = 0, \theta^t) \times P(S_{ig} = 0|Z_g = 1, \theta^t) \times P(Z_g = 1|\theta^t)\right)$$
$$+ w_{11,ig}^{t-1} \log\left(P(y_{ig}|S_{ig} = 1, \theta^t) \times P(S_{ig} = 1|Z_g = 1, \theta^t) \times P(Z_g = 1|\theta^t)\right)$$

$$Q(\theta^t|\theta^{t-1}) = \sum_{g=1}^{G}\sum_{i=1}^{N_g} w_{00,ig}^{t-1} \log\left(P_{null}(y_{ig}|\alpha^t) \times 1 \times (1 - q^t)\right) \tag{4.23}$$
$$+ w_{01,ig}^{t-1} \log\left(P_{alt}(y_{ig}|\alpha^t) \times 0 \times (1 - q^t)\right)$$
$$+ w_{10,ig}^{t-1} \log\left(P_{null}(y_{ig}|\alpha^t) \times (1 - p^t) \times q^t\right)$$
$$+ w_{11,ig}^{t-1} \log\left(P_{alt}(y_{ig}|\alpha^t) \times p^t \times q^t\right)$$

$$w_{00,ig}^{t-1} = \frac{P(Z_g = 0, S_{ig} = 0, y_{ig}|\theta^{t-1})}{P(y_{ig}|\theta^{t-1})} = \frac{P_{null}(y_{ig}|\alpha) \times 1 \times (1 - q)}{P_{null}(y_{ig}|\alpha) \times ((1 - q) + q \times (1 - p)) + P_{alt}(y_{ig}|\alpha) \times p \times q} \tag{4.24}$$

$$\boxed{w_{00,ig}^{t-1} = \frac{(1-q) \times P_{null}(y_{ig})}{(1-pq)P_{null}(y_{ig}) + pq \times P_{alt}(y_{ig})}} \tag{4.25}$$

$$\boxed{w_{01,ig}^{t-1} = 0} \tag{4.26}$$

$$\boxed{w_{10,ig}^{t-1} = \frac{(1-p) \times q \times P_{null}(y_{ig})}{(1-pq)P_{null}(y_{ig}) + pq \times P_{alt}(y_{ig})}} \tag{4.27}$$

$$\boxed{w_{11,ig}^{t-1} = \frac{pq \times P_{alt}(y_{ig})}{(1-pq)P_{null}(y_{ig}) + pq \times P_{alt}(y_{ig})}} \tag{4.28}$$

We take the derivate of Q function with respect to each parameter: $\frac{\partial Q}{\partial q} = 0$, $\frac{\partial Q}{\partial p} = 0$, $\frac{\partial Q}{\partial \alpha} = 0$

$$\begin{aligned} \frac{\partial Q}{\partial q} = \sum_{g=1}^{G}\sum_{i=1}^{N_g} & w_{00,ig}^{t-1} \frac{\partial}{\partial q}(\log(P_{null}(y_{ig}|\alpha) + \log(1-q)) \\ & + w_{10,ig}^{t-1} \frac{\partial}{\partial q}(\log(P_{null}(y_{ig}|\alpha) + \log(1-p) + \log q) \\ & + w_{11,ig}^{t-1} \frac{\partial}{\partial q}(\log(P_{alt}(y_{ig}|\alpha) + \log p + \log q) \end{aligned} \tag{4.29}$$

$$\frac{\partial Q}{\partial q} = \sum_{g=1}^{G}\sum_{i=1}^{N_g} w_{00,ig}^{t-1} \times (\frac{-1}{1-q}) + w_{10,ig}^{t-1} \times (\frac{1}{q}) + w_{11,ig}^{t-1} \times \frac{1}{q} = 0 \tag{4.30}$$

$$\boxed{q = \frac{\sum_{g=1}^{G}\sum_{i=1}^{N_g} w_{10,ig}^{t-1} + w_{11,ig}^{t-1}}{\sum_{g=1}^{G}\sum_{i=1}^{N_g} w_{00,ig}^{t-1} + w_{10,ig}^{t-1} + w_{11,ig}^{t-1}} = 1 - \frac{\sum_{g=1}^{G}\sum_{i=1}^{N_g} w_{00,ig}}{\sum_{g=1}^{G} N_g}} \tag{4.31}$$

$$\frac{\partial Q}{\partial p} = \sum_{g=1}^{G}\sum_{i=1}^{N_g} w_{00,ig}^{t-1} \frac{\partial}{\partial p}(\log(P_{null}(y_{ig}|\alpha) + \log(1-q))$$

$$+ w_{10,ig}^{t-1} \frac{\partial}{\partial p}(\log(P_{null}(y_{ig}|\alpha) + \log(1-p) + \log q) \tag{4.32}$$
$$+ w_{11,ig}^{t-1} \frac{\partial}{\partial p}(\log(P_{alt}(y_{ig}|\alpha) + \log p + \log q)$$

$$\frac{\partial Q}{\partial p} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{10,ig}^{t-1} \times (\frac{-1}{1-p}) + w_{11,ig}^{t-1} \times \frac{1}{p} = 0 \tag{4.33}$$

$$\boxed{p = \frac{\sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{11,ig}^{t-1}}{\sum_{g=1}^{G} \sum_{i=1}^{N_g} (w_{10,ig} + w_{11,ig})}} \tag{4.34}$$

$$\frac{\partial Q}{\partial \alpha} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{00,ig}^{t-1} \frac{\partial}{\partial \alpha}(\log(P_{null}(y_{ig}) + \log(1-q)) \tag{4.35}$$
$$+ w_{10,ig}^{t-1} \frac{\partial}{\partial \alpha}(\log(P_{null}(y_{ig}) + \log(1-p) + \log q)$$
$$+ w_{11,ig}^{t-1} \frac{\partial}{\partial \alpha}(\log(P_{alt}(y_{ig}|\alpha) + \log p + \log q)$$

$$\frac{\partial Q}{\partial \alpha} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{11,ig}^{t-1} \frac{\partial}{\partial \alpha}(\log \alpha + (\alpha - 1) \log y_{ig}) \tag{4.36}$$

$$\boxed{\alpha = -\frac{\sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{11,ig}^{t-1}}{\sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{11,ig}^{t-1} \times \log y_{ig}}} \tag{4.37}$$

We used the gradient decent algorithm with the gradient function computed as the following:

$$\nabla(Q) = \begin{bmatrix} \frac{\partial Q}{\partial \alpha} = & \sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{11,ig}^{t-1} (\frac{1}{\alpha} + \log(y_{ig})) \\ \frac{\partial Q}{\partial q} = & \sum_{g=1}^{G} \sum_{i=1}^{N_g} \frac{1}{q} (1 - \frac{w_{00,ig}^{t-1}}{1-q}) \\ \frac{\partial Q}{\partial p} = & \sum_{g=1}^{G} \sum_{i=1}^{N_g} (\frac{w_{11,ig}^{t-1}}{p} - \frac{w_{00,ig}^{t-1}}{1-p}) \end{bmatrix} \tag{4.38}$$

At each step, we update $\theta^t = \theta^{t-1} - \lambda \times \nabla(Q)$ where $\theta = (\alpha, q, p)$ and λ is learning rate.

$\alpha^t = \alpha^{t-1} - \lambda \times \frac{\partial Q}{\partial \alpha}$

$q^t = q^{t-1} - \lambda \times \frac{\partial Q}{\partial q}$

$p^t = p^{t-1} - \lambda \times \frac{\partial Q}{\partial p}$

4.3.5 Inference of the Hidden Parameters

The posterior probability of Z_g is the following:

$$\begin{aligned} P(Z_g = 1 | y_{ig}, \alpha, q, p) &= \prod_{i=1}^{N_g} \frac{P(Z_g = 1, y_{ig} | \alpha, q, p)}{P(y_{ig} | \alpha, q, p)} \\ &= \prod_{i=1}^{N_g} \frac{\sum_{k=0}^{1} P(S_{ig} = k | Z_g = 1) \times P(y_{ig} | S_{ig} = k)}{\sum_{k=0}^{1} \sum_{j=0}^{1} P(Z_g = j | q) \times P(S_{ig} = k | Z_g = j) \times P(y_{ig} | S_{ig} = k)} \end{aligned} \tag{4.39}$$

$$P(Z_g = 1 | y_{ig}, \alpha, q, p) = \prod_{i=1}^{N_g} \frac{(1-p) q P_{null}(y_{ig}) + pq P_{alt}(y_{ig} | \alpha)}{(1-pq) P_{null}(y_{ig}) + pq P_{alt}(y_{ig} | \alpha)} \tag{4.40}$$

$$\boxed{P(Z_g = 1 | y_{ig}, \alpha, q, p) = \prod_{i=1}^{N_g} (w_{10,ig} + w_{11,ig}) = \prod_{i=1}^{N_g} (1 - w_{00,ig})} \tag{4.41}$$

$$\begin{aligned} \frac{P(Z_g = 1 | D)}{P(Z_g = 0 | D)} &= \frac{\frac{P(Z_g=1, D)}{P(D)}}{\frac{P(Z_g=0, D)}{P(D)}} = \frac{P(D | Z_g = 1) \times P(Z_g = 1)}{P(D | Z_g = 0) \times P(Z_g = 0)} \\ &= \frac{q}{1-q} \prod_{i=1}^{N_g} \frac{P(D_i | Z_g = 1)}{P(D_i | Z_g = 0)} \geq 1 \end{aligned} \tag{4.42}$$

$$\frac{P(Z_g=1|D)}{P(Z_g=0|D)} = \frac{P(D|Z_g=1)\times P(Z_g=1)}{P(D|Z_g=0)\times P(Z_g=0)} = \frac{P(Z_g=1)}{P(Z_g=0)}\prod_{i=1}^{N_g}\frac{P(y_{ig}|Z_g=1)}{P(y_{ig}|Z_g=0)} = \tag{4.43}$$

$$\prod_{i=1}^{N_g}\frac{P(y_{ig}|S_{ig}=0)\times P(S_{ig}=0|Z_g=1)+P(y_{ig}|S_{ig}=1)\times P(S_{ig}=1|Z_g=1)}{P(y_{ig}|S_{ig}=0)\times P(S_{ig}=0|Z_g=0)+P(y_{ig}|S_{ig}=1)\times P(S_{ig}=1|Z_g=0)}\times\frac{q}{1-q}$$

$$\prod_{i=1}^{N_g}\frac{P_{null}(y_{ig})\times(1-p)+P_{alt}(y_{ig}|\alpha)\times p}{P_{null}(y_{ig})\times 1+P_{alt}(y_{ig}|\alpha)\times 0}\geq\frac{1-q}{q}$$

$$\boxed{Z_g=\begin{cases}1 & \text{if } \sum_{i=1}^{N_g}\log(1-p+\frac{p\times P_{alt}(y_{ig}|\alpha)}{null(y_{ig})})\geq\log(\frac{1-q}{q})\\ 0 & \text{Otherwise}\end{cases}} \tag{4.44}$$

$$\frac{P(S_{ig}=1|D)}{P(S_{ig}=0|D)} = \frac{P(D|S_{ig}=1)\times P(S_{ig}=1)}{P(D|S_{ig}=0)\times P(S_{ig}=0)} \tag{4.45}$$

The posterior probability of $S_{ig}=1$ is the following:

$$P(S_{ig}=1|y_{ig},\alpha,q,p) = \frac{P(S_{ig}=1,y_{ig}|\alpha,q,p)}{P(y_{ig}|\alpha,q,p)} = \frac{P(y_{ig}|S_{ig}=1,\alpha,q,p)\times P(S_{ig}=1|\alpha,p,q)}{\sum_{k=0}^{1}\sum_{j=0}^{1}P(y_{ig},S_ig=j,Z_g=k|\alpha,q,p)} \tag{4.46}$$

We first calculate $P(S_{ig}=1|\alpha,p,q)$:

$$P(S_{ig}=1|\alpha,p,q)=\sum_{j=0}^{1}P(Z_g=j|q)\times P(S_{ig}=1|Z_g=j) \tag{4.47}$$

Filling in:

$$P(S_{ig}=1)=p\times q+0\times(1-q)=pq \tag{4.48}$$

$$P(S_{ig}=0)=(1-p)\times q+1\times(1-q)=1-pq \tag{4.49}$$

$$
\begin{aligned}
P(S_{ig}=1|y_{ig},\alpha,q,p) &= \frac{P(S_{ig}=1,y_{ig}|\alpha,q,p)}{P(y_{ig}|\alpha,q,p)} \qquad (4.50)\\
&= \frac{\sum_{j=0}^{1} P(Z_g=j|q)\times P(S_{ig}=1|Z_g=j)\times P(Y_{ig}|S_{ig}=1)}{\sum_{k=0}^{1}\sum_{j=0}^{1} P(Z_g=j|q)\times P(S_{ig}=k|Z_g=j)\times P(Y_{ig}|S_{ig}=k)}\\
&= \frac{pqP_{alt}(y_{ig})}{pqP_{alt}(y_{ig}|\alpha)+(1-pq)P_{null}(y_{ig})} = w_{11,ig}
\end{aligned}
$$

$$
\boxed{P(S_{ig}=1|y_{ig},\alpha,q,p)=w_{11,ig}} \qquad (4.51)
$$

$$
\frac{P(S_{ig}=1|D)}{P(S_{ig}=0|D)} = \frac{P_{alt}(y_{ig}|\alpha)}{P_{null}(y_{ig})}\times\frac{pq}{1-pq} \qquad (4.52)
$$

$$
\boxed{S_{ig}=\begin{cases}1 & \text{if } \frac{P_{alt}(y_{ig}|\alpha)}{P_{null}(y_{ig})}\geq\frac{1}{pq}-1\\ 0 & \text{Otherwise}\end{cases}} \qquad (4.53)
$$

4.3.6 Log Likelihood Ratio

We estimate the parameters of the model using the above approach. Then, we compute the log likelihood of the data given the parameters. For the null model, we set parameter $p=0$ or $q=0$ or $\alpha=1$ and compute the differences in the log likelihood of the alternative model compared to the null model as the following.

$$
LLR = \log P(y|\mathbf{A},H=1) - \log P(y|\mathbf{A},H=0) \qquad (4.54)
$$

We scan 100 Kbp upstream and 100 Kbp downstream of the transcription start site for the gene g for to annotate binding sites for TF T. For each PWM for the TF T, we report the difference in the LLR score at the positions around the location of the eQTL SNPs; LLR of the reference and LLR of the alternative allele is computed using MOODS package in python [116]. We threshold the LLR scores at zero since negative LLR for the motif indicates that the TF has no binding sites at the location of the SNP. The ΔLLR for the motif is computed as the difference between LLR reference and LLR of the alternative allele. The motif score is reported for each TF binding site overlapping with the eQTL and we take the maximum ΔLLR score for each eQTL position. Since LLR depends on the length of the TFBS motif, we normalize the scores by ranking the ΔLLRs and computing a cumulative distribution function (CDF) based on the ranked values. For each eQTL position s, the p-value of the ΔLLR score is computed as $1-CDF(s)$. The two p-values, one for the eQTL and one

for the motif will be averaged to get y values. Then the PGM model is trained with the y and we get an estimate of the parameters of the PGM model fitting the data (y_{ig} for all SNPs i of the gene g. For each gene g, the direction of LLR change and the expression change (reference vs alternative) carries information on the role of the TF in the regulation of the gene. We compute the sign of $\Delta LLR \times \Delta Expr$ to infer the role of the TF. Here, $\Delta LLR = LLR_{ref} - LLR_{alt}$ and $\Delta Expr = Expr_{ref} - Expr_{alt}$ where $Expr_s$ is the average expression of the gene g for all individuals carrying SNP s.

In each tissue t, we estimate the likelihood of the PGM model using the trained parameters from the data computed above. We also estimate the likelihood of the PGM under the null model and compare the likelihood ratios for all PWM files of the TF T. There are multiple PWM files for some TFs. We chose the PWM with the best likelihood ratio to represent such TFs. Once we have the best PWM for all TFs, we order the TFs based on their best likelihood ratios and take the top 20 TFs in each tissue. The union set of the SNPs where there is an edge ($Z_g = 1$) between any of the 20 TFs and gene g forms the important gene set in the tissue t. We chose the functional SNPs from the SNPs that fall nearby the gene set that comes out of the previous analysis conditioned on $S_{ig} = 1$. Then we get a ranking of those SNPs based on the and the posterior $\frac{P(S_{ig}=1|D)}{P(S_{ig}=0|D)}$.

4.3.7 Tensorflow Implementation

We also have a TensorFlow implementation available for download, although this implementation was not used to generate the results of this study.

4.3.8 Code Location

Both the python code used to generate the results of this chapter and the TensorFlow code are available in a GitHub repository located at `https://github.com/khajoue2/pgm` and with a link to the GitHub repository at `veda.cs.uiuc.edu/pgm` .

4.4 USING THE PROBABILISTIC GRAPHICAL MODELING TO PRIORITIZE SNPS IN SIMULATED AND REAL DATA

4.4.1 Simulated Data

To evaluate the accuracy of our model, we applied the model on a simulated dataset. We generated synthetic data for $G = 500$ genes, where there are $N_g = 100$ SNPs per gene.

The data consists of eQTL p-values and motif score p-values for each SNP. These scores are generated from either a uniform distribution or a beta distribution with parameters $(\alpha = 0.2, \beta = 1)$ depending on the values of the hidden variables Z_g and S_{ig}.The hidden variables are fixed based on the following assumptions: 10% of the genes are targeted by the TF T $(q, 0.1)$ and for each g that is a target of the T, 5% of the SNPs are mediating the effect of T on expression of the gene g (parameter $p = 0.05$).

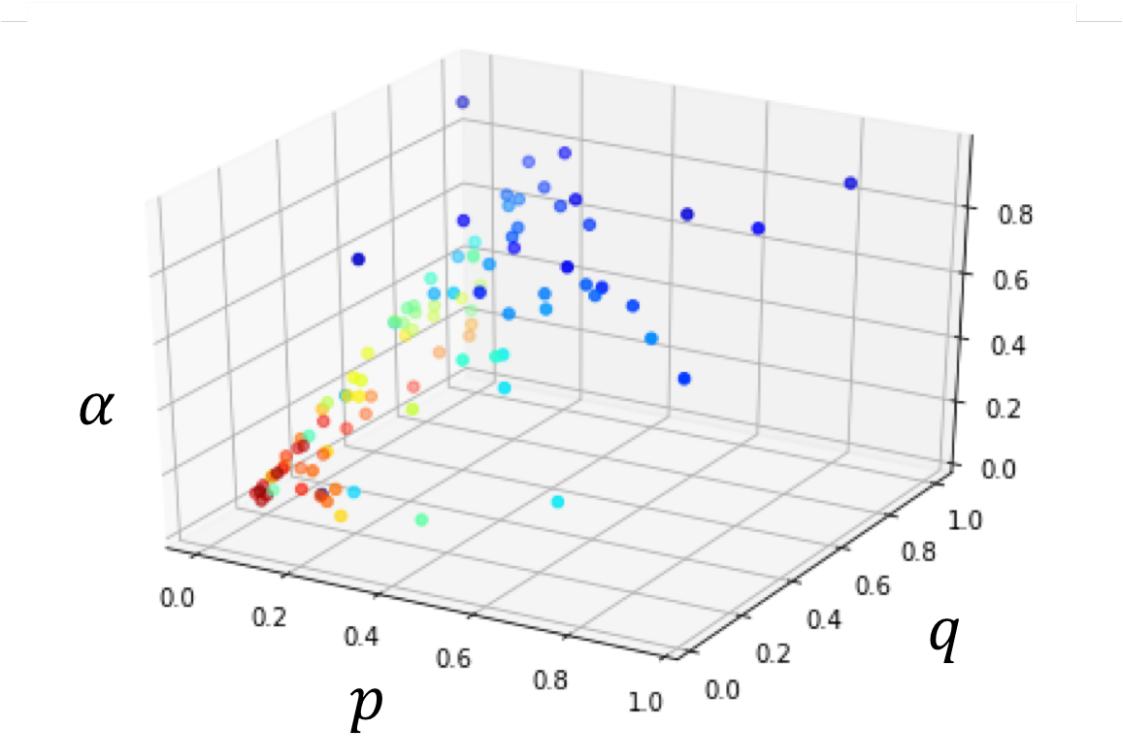

Figure 4.6: **Simulation Parameter Estimation** The EM algorithm converges to the true parameters of the simulated data. Each dot shows an iteration of the EM and the red colored points are the close points to the true parameters.

There are $G \times q = 50$ genes with $Z_g = 1$ and for each gene, $N_g \times p = 5$ SNPs with $S_{ig} = 1$ Given the simulated data, we estimate the parameters of the model and predict the hidden values. We report the accuracy of the parameter estimation using the EM algorithm by computing the true positive, false positive, true negative and false negative values. Figure 4.6 shows that the EM algorithm converges to the true parameters after several iterations. We tested the sensitivity of the PGM to the parameter estimations.

It is important to know that the EM algorithm does not always converge to the true parameters of the model since it is a local optimization method. To avoid that, we selected

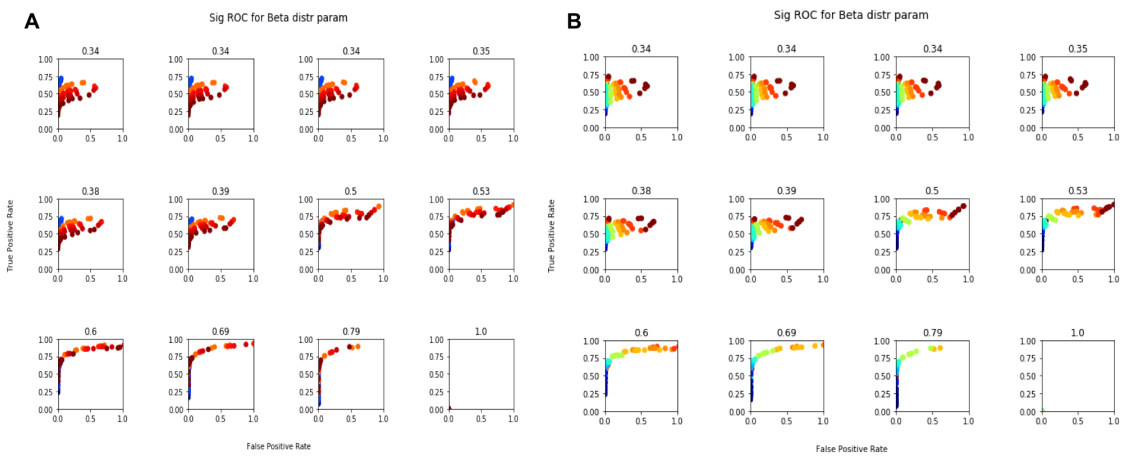

Figure 4.7: **Sensitivity of the model to the parameter estimation** Each panel shows the ROC plot for a fixed alpha on the title. The colors in (A) correspond to the p parameters and the colors in (B) corresponds to the q parameter.

1000 starting points for the EM algorithm and then the final converged parameters are compared against each other. We select the top 10 parameters converged from the EM and estimate an ensemble of models for the posterior probability of the SNPs. The posterior is computed as the average posteriors of these 10 parameters and we take a union over the SNPs that score highly from this ensemble.

4.4.2 Real Data

We processed the data for each tissue in the GTEx project separately (44 different tissues). The transcription factors modeled here were selected based on the presence of experimentally confirmed motifs for the TFs in CIS-BP dataset. For each transcription factor T in a tissue t, we trained the PGM and evaluated the likelihood of data under the alternative model and compared it to the likelihood of the data under the null model. GTEx set of eQTL are reported at the p-value significance around 10^{-4} and it does not provide the distribution of the whole data. To resolve that, we downloaded all association data from the GTEx project and included the QTL SNPs in our analysis. We only included the QTL SNPs for which the motif scores was computed; since the number of SNPs in the complete QTL data is significantly larger and majority of these SNPs are not important in the final analysis.

For each TF T, we selected the motif with the best likelihood score of the data. Then we sorted all TFs based on their likelihood ratios (alternative vs null) and selected the top 20 TFs. Then, we selected the genes that are the targets of these TFs based on the posterior probabilities computed in equation 4.3.5 and 4.44. For every gene g that passes the threshold

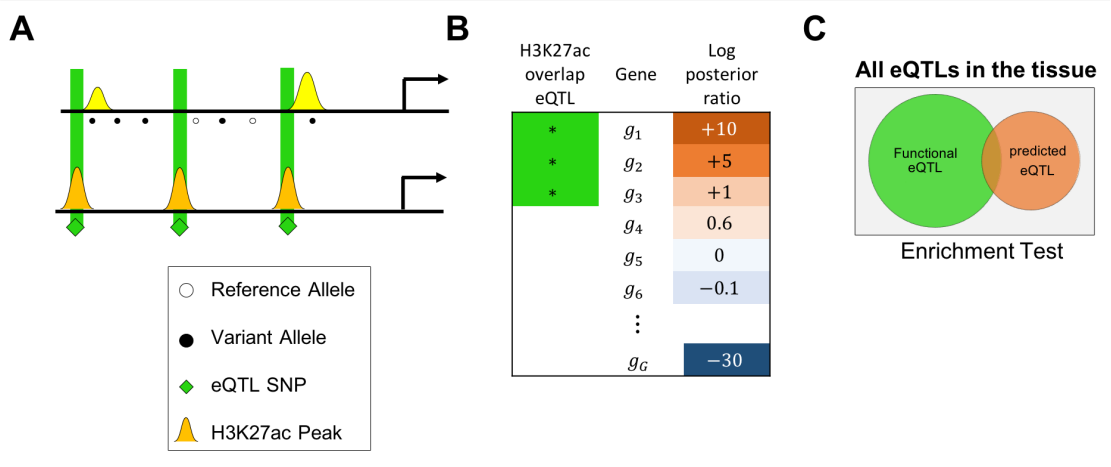

Figure 4.8: **Validations** A) We define the functional set of SNPs as the eQTL SNPs that overlap with the H3K27ac marks. B) PGM model provides a ranking of the SNPs based on the posterior probability ratios. We select the baseline as the eQTL SNPs ranked by their p-values restricting one eQTL per gene. C) Hyper-geometric test of the significance

on $\frac{P(Z_g=1|Data)}{P(Z_g=0|Data)}$, we select the nearby SNP i that is associated with the gene g based on the largest posterior $\frac{P(S_{ig}=1|Data)}{P(S_{ig}=0|Data)}$. We get one SNP per gene from this ranking. Note that it is possible that the same SNP ranks highest for two different genes since the SNP has large motif score and the eQTL score was also large for more than one gene. The baseline method to compare our results with is the following. We select one SNP per gene in the tissue based on the eQTL p-value. Then we sort these SNPs from the strongest p-value to the smallest. Then we take the top k SNPs from both this best eQTL method (baseline) and the PGM method sorted by the posterior probability of the SNP (our method). We set k to be the top 100 and top 1000 motifs.

We define a functional set of SNPs based on the overlap of the eQTL SNPs from GTEx data and the H3K27ac marks from the ENCODE data [10] for the same tissue (figure 4.8).The expected number of SNPs is calculated as $k \times \frac{\#functionaleQTLs}{\#alleQTLs}$. The predicted SNPs overlap with the functional set of SNPs shows the power of our method compared to the baseline.

For a second validation, we defined the functional SNPs as the set of the eQTL SNPs that overlap with the eRNA data in different cancer [117]. Then, we conducted a similar enrichment test and reported the hyper-geometric p-values. We summarized the results in the following tables for four different tissues (table A.4, A.5, A.6 and A.7). The PGM method outperforms the baseline with the H3K27ac broad peaks. This chromatin mark shows the active enhancer regions and it is expected to have multiple TF binding sites. The H3K27ac narrow peaks does not show an improvement over the baseline but this is expected since the narrow peaks mark the regions that are not necessarily bound by the TFs since

the histones would occupy the DNA at the narrow peak locations. The number of overlaps with the narrow peak in the baseline is consistent with the ratio of the width of the peak from broad to narrow.

4.5 DISCUSSIONS

Ideally, we would like to link the whole genome sequence and use all variations to infer such relations. We may report a SNP as causal in a disease, but it might be the case that the actual causal variation could be another genetically linked but un-probed variant such as an indel or a copy number variant. This means that those variants may be associated with a different TF or a different gene. Another limitation in our study is using position weight matrix of only 700 TFs (out of a total of 2000 known human TFs) since majority of TFs are still devoid of DNA binding specificity. We can improve our result by a probabilistic model that finds the SNPs that mediate the effect of the TF on the gene and can explain gene expression variation across individuals. Using a probabilistic graphical model, we can then compute the posterior probability that the variant mediates the TF's influence on expression. This model allows us to infer the role of the TF as an activator or a repressor. Furthermore, this will allow us to use the information in the binding cite combined with the gene expression sign (loss vs. gain of binding strength), not just its magnitude. We believe that this approach allows us to understand the regulatory impact of a variant based on the context of the cell and it can also be extended to prioritization of GWAS SNPs associated with phenotypic variation.

CHAPTER 5: REFERENCES

[1] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 2010.

[2] R. R. Haraksingh and M. P. Snyder, "Impacts of variation in the human genome on gene regulation," *Journal of molecular biology*, vol. 425, no. 21, pp. 3970–3977, 2013.

[3] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung, "Genetic analysis of genome-wide variation in human gene expression," *Nature*, vol. 430, no. 7001, p. 743, 2004.

[4] D. S. Paul, N. Soranzo, and S. Beck, "Functional interpretation of non-coding sequence variation: concepts and challenges," *Bioessays*, vol. 36, no. 2, pp. 191–199, 2014.

[5] F. Zhang and J. R. Lupski, "Non-coding genetic variants in human disease," *Human molecular genetics*, vol. 24, no. R1, pp. R102–R110, 2015.

[6] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody et al., "Systematic localization of common disease-associated variation in regulatory dna," *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[7] L. D. Ward and M. Kellis, "Interpreting noncoding genetic variation in complex traits and human disease," *Nature biotechnology*, vol. 30, no. 11, p. 1095, 2012.

[8] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional regulatory elements in the human genome," *Annu. Rev. Genomics Hum. Genet.*, vol. 7, pp. 29–59, 2006.

[9] Z. Duren, X. Chen, R. Jiang, Y. Wang, and W. H. Wong, "Modeling gene regulation from paired expression and chromatin accessibility data," *Proceedings of the National Academy of Sciences*, vol. 114, no. 25, pp. E4914–E4923, 2017.

[10] E. P. Consortium et al., "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, p. 57, 2012.

[11] A. Korte and A. Farlow, "The advantages and limitations of trait analysis with gwas: a review," *Plant methods*, vol. 9, no. 1, p. 29, 2013.

[12] Y. G. Tak and P. J. Farnham, "Making sense of gwas: using epigenomics and genome engineering to understand the functional relevance of snps in non-coding regions of the human genome," *Epigenetics & chromatin*, vol. 8, no. 1, p. 57, 2015.

[13] K. Wetterstrand, "Dna sequencing costs: Data from the nhgri genome sequencing program(gsp) available at: www. genome. gov/sequencingcostsdata. accessed 06-2018. of genomes using long-read sequencing technology,"," *Genome Res*, vol. 24, pp. 688–696, 2014.

[14] G. M. Cooper, R. E. Hausman, and R. E. Hausman, *The cell: a molecular approach*. ASM press Washington, DC, 2000, vol. 10.

[15] G. E. Moore, M. Ishida, C. Demetriou, L. Al-Olabi, L. J. Leon, A. C. Thomas, S. Abu-Amero, J. M. Frost, J. L. Stafford, Y. Chaoqun et al., "The role and interaction of imprinted genes in human fetal growth," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1663, p. 20140074, 2015.

[16] M. Garieri, O. Delaneau, F. Santoni, R. J. Fish, D. Mull, P. Carninci, E. T. Dermitzakis, S. E. Antonarakis, and A. Fort, "The effect of genetic variation on promoter usage and enhancer activity," *Nature communications*, vol. 8, no. 1, p. 1358, 2017.

[17] W. Guo, L. Zhu, S. Deng, X. Zhao, and D. Huang, "Understanding tissue-specificity with human tissue-specific regulatory networks," *Science China Information Sciences*, vol. 59, no. 7, p. 070105, 2016.

[18] R. Sayal, J. M. Dresch, I. Pushel, B. R. Taylor, and D. N. Arnosti, "Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early drosophila embryo," *Elife*, vol. 5, p. e08445, 2016.

[19] A. Barbeira, S. P. Dickinson, J. M. Torres, E. S. Torstenson, J. Zheng, H. E. Wheeler, K. P. Shah, T. Edwards, D. Nicolae, N. J. Cox et al., "Integrating tissue specific mechanisms into gwas summary results," *BioRxiv*, p. 045260, 2017.

[20] M. A. H. Samee, B. Lim, N. Samper, H. Lu, C. A. Rushlow, G. Jiménez, S. Y. Shvartsman, and S. Sinha, "A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data," *Cell systems*, vol. 1, no. 6, pp. 396–407, 2015.

[21] V. V. Gursky, K. N. Kozlov, I. V. Kulakovskiy, A. Zubair, P. Marjoram, D. S. Lawrie, S. V. Nuzhdin, and M. G. Samsonova, "Translating natural genetic variation to gene expression in a computational model of the drosophila gap gene regulatory network," *PloS one*, vol. 12, no. 9, p. e0184657, 2017.

[22] S. Bonn, R. P. Zinzen, C. Girardot, E. H. Gustafson, A. Perez-Gonzalez, N. Delhomme, Y. Ghavi-Helm, B. Wilczyński, A. Riddell, and E. E. Furlong, "Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development," *Nature genetics*, vol. 44, no. 2, p. 148, 2012.

[23] C. Blatti, M. Kazemian, S. Wolfe, M. Brodsky, and S. Sinha, "Integrating motif, dna accessibility and gene expression data to build regulatory maps in an organism," *Nucleic acids research*, vol. 43, no. 8, pp. 3998–4012, 2015.

[24] D. Svetlichnyy, H. Imrichova, M. Fiers, Z. K. Atak, and S. Aerts, "Identification of high-impact cis-regulatory mutations using transcription factor specific random forest models," *PLoS computational biology*, vol. 11, no. 11, p. e1004590, 2015.

[25] M. Kazemian, C. Blatti, A. Richards, M. McCutchan, N. Wakabayashi-Ito, A. S. Hammonds, S. E. Celniker, S. Kumar, S. A. Wolfe, M. H. Brodsky et al., "Quantitative analysis of the drosophila segmentation regulatory network using pattern generating potentials," *PLoS biology*, vol. 8, no. 8, p. e1000456, 2010.

[26] M. Spivakov, J. Akhtar, P. Kheradpour, K. Beal, C. Girardot, G. Koscielny, J. Herrero, M. Kellis, E. E. Furlong, and E. Birney, "Analysis of variation at transcription factor binding sites in drosophila and humans," *Genome biology*, vol. 13, no. 9, p. R49, 2012.

[27] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, p. 931, 2015.

[28] D. Lee, D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, and M. A. Beer, "A method to predict the impact of regulatory variants from dna sequence," *Nature genetics*, vol. 47, no. 8, p. 955, 2015.

[29] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks," in *Biocomputing 2001*. World Scientific, 2000, pp. 422–433.

[30] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul, "Predicting expression patterns from regulatory sequence in drosophila segmentation," *Nature*, vol. 451, no. 7178, p. 535, 2008.

[31] H. Janssens, S. Hou, J. Jaeger, A.-R. Kim, E. Myasnikova, D. Sharp, and J. Reinitz, "Quantitative and predictive model of transcriptional control of the drosophila melanogaster even skipped gene," *Nature genetics*, vol. 38, no. 10, p. 1159, 2006.

[32] R. P. Zinzen, K. Senger, M. Levine, and D. Papatsenko, "Computational models for neurogenic gene expression in the drosophila embryo," *Current Biology*, vol. 16, no. 13, pp. 1358–1365, 2006.

[33] M. A. White, D. S. Parker, S. Barolo, and B. A. Cohen, "A model of spatially restricted transcription in opposing gradients of activators and repressors," *Molecular systems biology*, vol. 8, no. 1, 2012.

[34] T. Ahsendorf, F. Wong, R. Eils, and J. Gunawardena, "A framework for modelling gene regulation which accommodates non-equilibrium mechanisms," *BMC biology*, vol. 12, no. 1, p. 102, 2014.

[35] J. Gertz, E. D. Siggia, and B. A. Cohen, "Analysis of combinatorial cis-regulation in synthetic and genomic promoters," *Nature*, vol. 457, no. 7226, p. 215, 2009.

[36] X. He, M. A. H. Samee, C. Blatti, and S. Sinha, "Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression," *PLoS computational biology*, vol. 6, no. 9, p. e1000935, 2010.

[37] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield et al., "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, no. 7527, p. 402, 2014.

[38] W. Li, D. Notani, and M. G. Rosenfeld, "Enhancers as non-coding rna transcription units: recent insights and future perspectives," *Nature Reviews Genetics*, vol. 17, no. 4, p. 207, 2016.

[39] A. Visel, E. M. Rubin, and L. A. Pennacchio, "Genomic views of distant-acting enhancers," *Nature*, vol. 461, no. 7261, p. 199, 2009.

[40] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban et al., "Variation in transcription factor binding among humans," *science*, vol. 328, no. 5975, pp. 232–235, 2010.

[41] D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad, "The functional consequences of variation in transcription factor binding," *PLoS genetics*, vol. 10, no. 3, p. e1004226, 2014.

[42] F. Khajouei and S. Sinha, "An information theoretic treatment of sequence-to-expression modeling," *PLoS computational biology*, vol. 14, no. 9, p. e1006459, 2018.

[43] F. Spitz and E. E. Furlong, "Transcription factors: from enhancer binding to developmental control," *Nature reviews genetics*, vol. 13, no. 9, p. 613, 2012.

[44] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, "Universally sloppy parameter sensitivities in systems biology models," *PLoS computational biology*, vol. 3, no. 10, p. e189, 2007.

[45] K. S. Brown and J. P. Sethna, "Statistical mechanical approaches to models with many poorly known parameters," *Physical review E*, vol. 68, no. 2, p. 021904, 2003.

[46] C. J. Wu and M. S. Hamada, *Experiments: planning, analysis, and optimization*. John Wiley & Sons, 2011, vol. 552.

[47] C. Kreutz and J. Timmer, "Systems biology: experimental design," *The FEBS journal*, vol. 276, no. 4, pp. 923–942, 2009.

[48] P. Flaherty, A. Arkin, and M. I. Jordan, "Robust design of biological experiments," in *Advances in neural information processing systems*, 2006, pp. 363–370.

[49] Y. Suleimenov, A. Ay, M. A. H. Samee, J. M. Dresch, S. Sinha, and D. N. Arnosti, "Global parameter estimation for thermodynamic models of transcriptional regulation," *Methods*, vol. 62, no. 1, pp. 99–108, 2013.

[50] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, vol. 90, no. 430, pp. 577–588, 1995.

[51] J. B. Weiss, T. Von Ohlen, D. M. Mellerick, G. Dressler, C. Q. Doe, and M. P. Scott, "Dorsoventral patterning in the drosophila central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity," *Genes & development*, vol. 12, no. 22, pp. 3591–3602, 1998.

[52] A. Stathopoulos and M. Levine, "Localized repressors delineate the neurogenic ectoderm in the early drosophila embryo," *Developmental biology*, vol. 280, no. 2, pp. 482–493, 2005.

[53] T. Von Ohlen and C. Q. Doe, "Convergence of dorsal, dpp, and egfr signaling pathways subdivides the drosophila neuroectoderm into three dorsal-ventral columns," *Developmental biology*, vol. 224, no. 2, pp. 362–372, 2000.

[54] C.-Y. Nien, H.-L. Liang, S. Butcher, Y. Sun, S. Fu, T. Gocha, N. Kirov, J. R. Manak, and C. Rushlow, "Temporal coordination of gene networks by zelda in the early drosophila embryo," *PLoS genetics*, vol. 7, no. 10, p. e1002339, 2011.

[55] M. Garcia and A. Stathopoulos, "Lateral gene expression in drosophila early embryos is supported by grainyhead-mediated activation and tiers of dorsally-localized repression," *PloS one*, vol. 6, no. 12, p. e29172, 2011.

[56] J. A. McDonald, S. Holbrook, T. Isshiki, J. Weiss, C. Q. Doe, and D. M. Mellerick, "Dorsoventral patterning in the drosophila central nervous system: the vnd homeobox gene specifies ventral column identity," *Genes & development*, vol. 12, no. 22, pp. 3603–3612, 1998.

[57] B. Lim, N. Samper, H. Lu, C. Rushlow, G. Jiménez, and S. Y. Shvartsman, "Kinetics of gene derepression by erk signaling," *Proceedings of the National Academy of Sciences*, vol. 110, no. 25, pp. 10 330–10 335, 2013.

[58] Y. Kasai, S. Stahl, and S. Crews, "Specification of the drosophila cns midline cell lineage: direct control of single-minded transcription by dorsal/ventral patterning genes," *Gene Expression, The Journal of Liver Research*, vol. 7, no. 3, pp. 171–189, 1998.

[59] J. B. Thomas, S. T. Crews, and C. S. Goodman, "Molecular genetics of the single-minded locus: a gene involved in the development of the drosophila nervous system," *Cell*, vol. 52, no. 1, pp. 133–141, 1988.

[60] J. R. Nambu, R. G. Franks, S. Hu, and S. T. Crews, "The single-minded gene of drosophila is required for the expression of genes important for the development of cns midline cells," *Cell*, vol. 63, no. 1, pp. 63–75, 1990.

[61] S. T. Crews, *PAS proteins: regulators and sensors of development and physiology*. Springer Science & Business Media, 2003.

[62] J. R. Nambu, J. O. Lewis, K. A. Wharton Jr, and S. T. Crews, "The drosophila single-minded gene encodes a helix-loop-helix protein that acts as a master regulator of cns midline development," *Cell*, vol. 67, no. 6, pp. 1157–1167, 1991.

[63] S. T. Crews, "Control of cell lineage-specific development and transcription by bhlh–pas proteins," *Genes & development*, vol. 12, no. 5, pp. 607–620, 1998.

[64] V. Morel and F. Schweisguth, "Repression by suppressor of hairless and activation by notch are required to define a single row of single-minded expressing cells in the drosophila embryo," *Genes & development*, vol. 14, no. 3, pp. 377–388, 2000.

[65] J. Cowden and M. Levine, "The snail repressor positions notch signaling in the drosophila embryo," *Development*, vol. 129, no. 7, pp. 1785–1793, 2002.

[66] M. D. Martín-Bermudo, A. Carmena, and F. Jiménez, "Neurogenic genes control gene expression at the transcriptional level in early neurogenesis and in mesectoderm specification," *Development*, vol. 121, no. 1, pp. 219–224, 1995.

[67] V. Morel, R. Le Borgne, and F. Schweisguth, "Snail is required for delta endocytosis and notch-dependent activation of single-minded expression," *Development genes and evolution*, vol. 213, no. 2, pp. 65–72, 2003.

[68] S. Bray and M. Furriols, "Notch pathway: making sense of suppressor of hairless," *Current Biology*, vol. 11, no. 6, pp. R217–R221, 2001.

[69] K.-W. Park and J.-W. Hong, "Mesodermal repression of single-minded in drosophila embryo is mediated by a cluster of snail-binding sites proximal to the early promoter," *BMB reports*, vol. 45, no. 10, pp. 577–582, 2012.

[70] C. Fraley and A. E. Raftery, "Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust," *Journal of Classification*, vol. 20, no. 2, pp. 263–286, 2003.

[71] C. Fraley and A. E. Raftery, "Mclust version 3 for r: Normal mixture modeling and model-based clustering," Citeseer, Tech. Rep., 2006.

[72] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[73] L. J. Zhu, R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe et al., "Flyfactorsurvey: a database of drosophila transcription factor binding specificities determined using the bacterial one-hybrid system," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D111–D117, 2010.

[74] A. Ay and D. N. Arnosti, "Mathematical modeling of gene expression: a guide for the perplexed biologist," *Critical reviews in biochemistry and molecular biology*, vol. 46, no. 2, pp. 137–151, 2011.

[75] J. M. Hernández-Lobato, M. A. Gelbart, M. W. Hoffman, R. P. Adams, and Z. Ghahramani, "Predictive entropy search for bayesian optimization with unknown constraints," *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015.

[76] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[77] D. V. Lindley et al., "On a measure of the information provided by an experiment," *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 986–1005, 1956.

[78] T. Kaplan, X.-Y. Li, P. J. Sabo, S. Thomas, J. A. Stamatoyannopoulos, M. D. Biggin, and M. B. Eisen, "Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early drosophila development," *PLoS genetics*, vol. 7, no. 2, p. e1001290, 2011.

[79] J. O. Yáñez-Cuna, E. Z. Kvon, and A. Stark, "Deciphering the transcriptional cis-regulatory code," *Trends in Genetics*, vol. 29, no. 1, pp. 11–22, 2013.

[80] B. Deplancke, D. Alpern, and V. Gardeux, "The genetics of transcription factor dna binding variation," *Cell*, vol. 166, no. 3, pp. 538–554, 2016.

[81] J. Bischof, M. Björklund, E. Furger, C. Schertel, J. Taipale, and K. Basler, "A versatile platform for creating a comprehensive uas-orfeome library in drosophila," *Development*, vol. 140, no. 11, pp. 2434–2442, 2013.

[82] J. Bischof, R. K. Maeda, M. Hediger, F. Karch, and K. Basler, "An optimized transgenesis system for drosophila using germ-line-specific φc31 integrases," *Proceedings of the National Academy of Sciences*, vol. 104, no. 9, pp. 3312–3317, 2007.

[83] W. Huang, A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia, A. M. Tarone, L. Turlapati, T. Zichner, D. Zhu, R. F. Lyman et al., "Natural variation in genome architecture among 205 drosophila melanogaster genetic reference panel lines," *Genome research*, vol. 24, no. 7, pp. 1193–1208, 2014.

[84] J. Phinchongsakuldit, S. MacArthur, and J. F. Brookfield, "Evolution of developmental genes: molecular microevolution of enhancer sequences at the ubx locus in drosophila and its impact on developmental phenotypes," *Molecular biology and evolution*, vol. 21, no. 2, pp. 348–363, 2004.

[85] J. Zeitlinger, R. P. Zinzen, A. Stark, M. Kellis, H. Zhang, R. A. Young, and M. Levine, "Whole-genome chip–chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo," *Genes & development*, vol. 21, no. 4, pp. 385–390, 2007.

[86] T. Werner, A. Hammer, M. Wahlbuhl, M. R. Bösl, and M. Wegner, "Multiple conserved regulatory elements with overlapping functions determine sox10 expression in mouse embryogenesis," *Nucleic acids research*, vol. 35, no. 19, pp. 6526–6538, 2007.

[87] J.-W. Hong, D. A. Hendrix, and M. S. Levine, "Shadow enhancers as a source of evolutionary novelty," *Science*, vol. 321, no. 5894, pp. 1314–1314, 2008.

[88] S. Horn, A. Figl, P. S. Rachakonda, C. Fischer, A. Sucker, A. Gast, S. Kadel, I. Moll, E. Nagore, K. Hemminki et al., "Tert promoter mutations in familial and sporadic melanoma," *Science*, vol. 339, no. 6122, pp. 959–961, 2013.

[89] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, "Mapping complex disease traits with global gene expression," *Nature Reviews Genetics*, vol. 10, no. 3, p. 184, 2009.

[90] D. C. Croteau-Chonka, A. J. Rogers, T. Raj, M. J. McGeachie, W. Qiu, J. P. Ziniti, B. J. Stubbs, L. Liang, F. D. Martinez, R. C. Strunk et al., "Expression quantitative trait loci information improves predictive modeling of disease relevance of non-coding genetic variation," *PloS one*, vol. 10, no. 10, p. e0140758, 2015.

[91] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: from polygenic to omnigenic," *Cell*, vol. 169, no. 7, pp. 1177–1186, 2017.

[92] O. Wagih, D. Merico, A. Delong, and B. J. Frey, "Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors," *BioRxiv*, p. 253427, 2018.

[93] I. Mogno, J. C. Kwasnieski, and B. A. Cohen, "Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants," *Genome research*, vol. 23, no. 11, pp. 1908–1915, 2013.

[94] X. Xie, C. Hanson, and S. Sinha, "Mechanistic interpretation of non-coding variants for discovering transcriptional regulators of drug response," *BMC biology*, vol. 17, no. 1, p. 62, 2019.

[95] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher et al., "Integration of summary data from gwas and eqtl studies predicts complex trait gene targets," *Nature genetics*, vol. 48, no. 5, p. 481, 2016.

[96] Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts et al., "Variations in dna elucidate molecular networks that cause disease," *Nature*, vol. 452, no. 7186, pp. 429–435, 2008.

[97] R. B. Williams, E. K. Chan, M. J. Cowley, and P. F. Little, "The influence of genetic variation on gene expression," *Genome research*, vol. 17, no. 12, pp. 1707–1716, 2007.

[98] W. Y. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd, "Genome-wide association studies: theoretical and practical concerns," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 109–118, 2005.

[99] E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 415–425, 2010.

[100] D. G. Hernandez, M. A. Nalls, M. Moore, S. Chong, A. Dillman, D. Trabzuni, J. R. Gibbs, M. Ryten, S. Arepalli, M. E. Weale et al., "Integration of gwas snps and tissue specific expression profiling reveal discrete eqtls for human traits in blood and brain," *Neurobiology of disease*, vol. 47, no. 1, pp. 20–28, 2012.

[101] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox, "Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas," *PLoS genetics*, vol. 6, no. 4, 2010.

[102] A. C. Nica, S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, and E. T. Dermitzakis, "Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations," *PLoS Genet*, vol. 6, no. 4, p. e1000895, 2010.

[103] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano, "Reverse engineering cellular networks," *Nature protocols*, vol. 1, no. 2, p. 662, 2006.

[104] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[105] J. Zhang, K. Lu, Y. Xiang, M. Islam, S. Kotian, Z. Kais, C. Lee, M. Arora, H.-w. Liu, J. D. Parvin et al., "Weighted frequent gene co-expression network mining to identify genes involved in genome stability," *PLoS computational biology*, vol. 8, no. 8, p. e1002656, 2012.

[106] P. Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek, "Addressing confounding artifacts in reconstruction of gene co-expression networks," *Genome biology*, vol. 20, no. 1, pp. 1–6, 2019.

[107] A. F. Siahpirani, D. Chasman, and S. Roy, "Integrative approaches for inference of genome-scale gene regulatory networks," in *Gene Regulatory Networks*. Springer, 2019, pp. 161–194.

[108] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young et al., "The genotype-tissue expression (gtex) project," *Nature genetics*, vol. 45, no. 6, p. 580, 2013.

[109] I. S. Kathiriya, E. P. Nora, and B. G. Bruneau, "Investigating the transcriptional control of cardiovascular development," *Circulation research*, vol. 116, no. 4, pp. 700–714, 2015.

[110] A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei et al., "Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals," *Genome research*, vol. 24, no. 1, pp. 14–24, 2014.

[111] M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook et al., "Determination and inference of eukaryotic transcription factor sequence specificity," *Cell*, vol. 158, no. 6, pp. 1431–1443, 2014.

[112] G. Consortium et al., "The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, 2015.

[113] G. Consortium et al., "Genetic effects on gene expression across human tissues," *Nature*, vol. 550, no. 7675, pp. 204–213, 2017.

[114] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution," *Nature Reviews Genetics*, vol. 10, no. 4, p. 252, 2009.

[115] A. C. Nica and E. T. Dermitzakis, "Expression quantitative trait loci: present and future," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1620, p. 20120362, 2013.

[116] J. H. Korhonen, K. Palin, J. Taipale, and E. Ukkonen, "Fast motif matching revisited: high-order pwms, snps and indels," *Bioinformatics*, vol. 33, no. 4, pp. 514–521, 2017.

[117] Z. Zhang, J.-H. Lee, H. Ruan, Y. Ye, J. Krakowiak, Q. Hu, Y. Xiang, J. Gong, B. Zhou, L. Wang et al., "Transcriptional landscape and clinical utility of enhancer rnas for erna-targeted therapy in cancer," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

APPENDIX A: SUPPLEMENTARY FIGURES AND TABLES

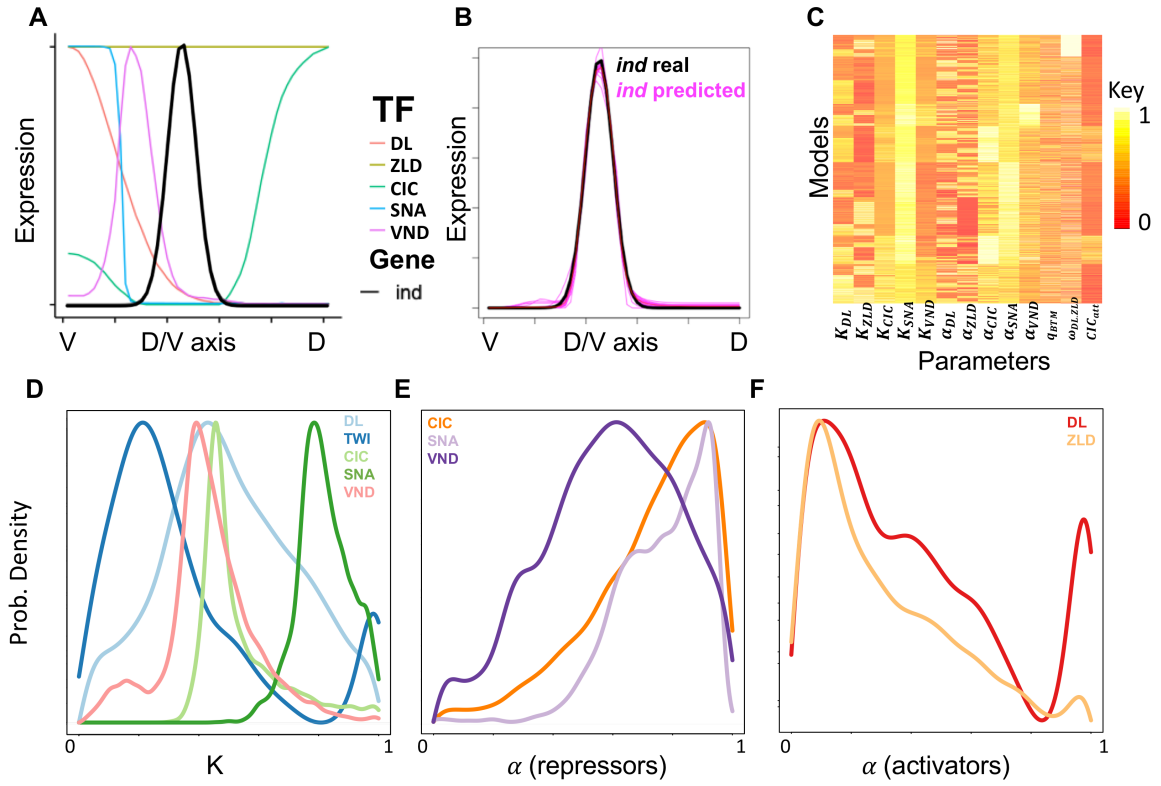

Figure A.1: The predicted ind expression from a wild-type ensemble of models and the distribution of parameters in the ensemble. (A) The expression domain for TFs and the 'ind' gene is shown along the Dorsal-ventral domain. The x-axis represents ventral (left) to dorsal (right) end of the D/V axis and the y-axis is the expression value from no expression to the maximum observed expression for each gene or TD, on a scale of 0 to 1. (B) Predicted average ind expression (magenta) from all models optimized to fit wild-type data (black). Each pink line shows the prediction of a single model in the ensemble (C) Each row is a model in the ensemble and each column corresponds to a parameter for the model. Each parameter is scaled to the range of 0 to 1. The K parameter for all TFs and α parameter of repressors are in logarithmic scale and the α parameter of activators, cooperativity and q_{BTM} are in linear scale. (D-F) Marginal densities of parameters of the ensemble. Each parameter vector is scaled to be in the same range. The x-axes in (D) and (E) are in logarithmic scale and in F in linear scale.

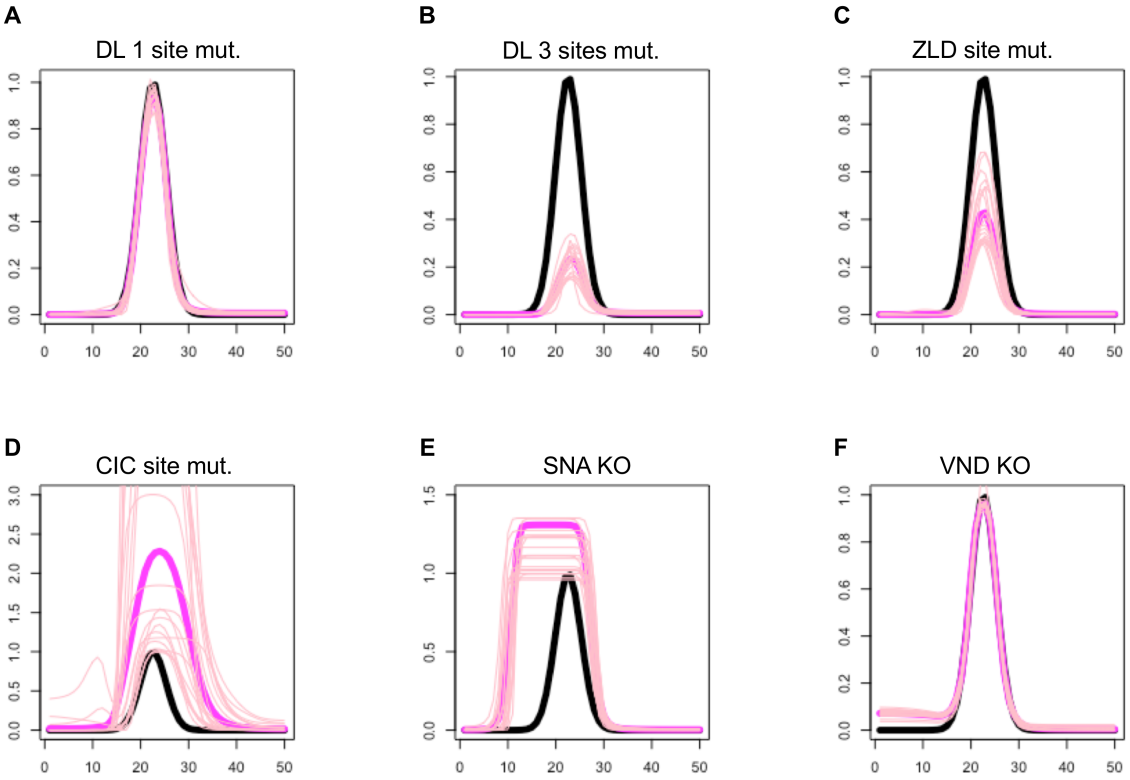

Figure A.2: Filtered ensemble for each of six perturbation experiments in the literature. The wild-type ensemble was filtered to retain only models that correctly predict the observed effects of an additional perturbation experiment (A.1 Table), thus yielding a smaller, filtered ensemble for each experiment. In each panel, wild-type ind expression is shown in black, pink curves represent predictions of models in the ensemble and magenta represents a weighted average of the ensemble predictions. (A) No change is observed in the expression when the strongest DL site is mutated. (B) Peak ind expression is reduced by 65% after 3 DL sites are mutated. (C) Peak ind expression is reduced by half upon mutations in ZLD binding sites. (D) ind expression expands dorsally when two sites of CIC is mutated. (E) ind expression expands ventrally in VND knockout. (F) The expression of ind is not changed in SNA knockout experiment.

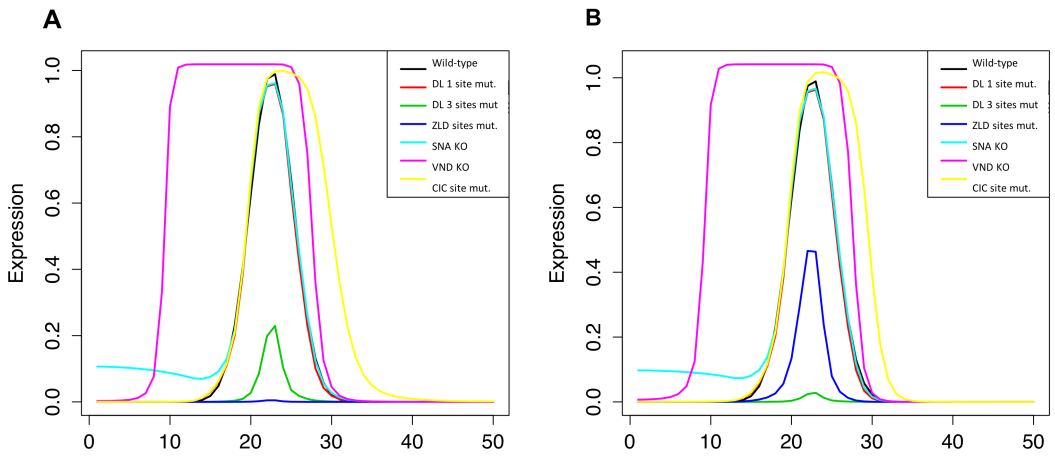

Figure A.3: Selecting a 'synthetic real' model. We searched for a model that not only has a good fit to the wild-type ind expression but also produces the known effects of perturbation experiments reported in the literature (A.1 Table). Such a model would then be used as the 'truth' for predicting the effects of other 'experiments'. Starting from the wild-type ensemble, we filtered models that predict the effect of CIC site mutation ('CIC site Mut.'), VND knockout ('VND KO'), SNA knockout ('SNA KO') and mutagenesis of the strongest predicted DL site ('DL 1 site Mut.') correctly, resulting in an ensemble of a few hundred models. Then, we checked the ability of these models to reproduce the effect of an experiment where three overlapping DL sites were mutagenized ('DL 3 site Mut.') and another experiment where the four strongest ZLD sites were mutagenized ('ZLD site mutation'). We were unable to find any model that could reproduce both results correctly. Thus, we used only one of these two filters to obtain an ensemble of models that can predict five out of six perturbation experiments correctly. Shown are the predictions for the wild-type condition and the six perturbation conditions, made by two distinct models, both of which fit wild-type data and perturbation experiments 'CIC site Mut.', 'VND KO', 'SNA KO', 'DL 1 site Mut.' as well as either (A) 'DL 3 site Mut.' (this model fails to reproduce the effect of 'ZLD site Mut.') or (B) 'ZLD site Mut.' (this model is unable to reproduce the effect of 'DL 3 site Mut.').

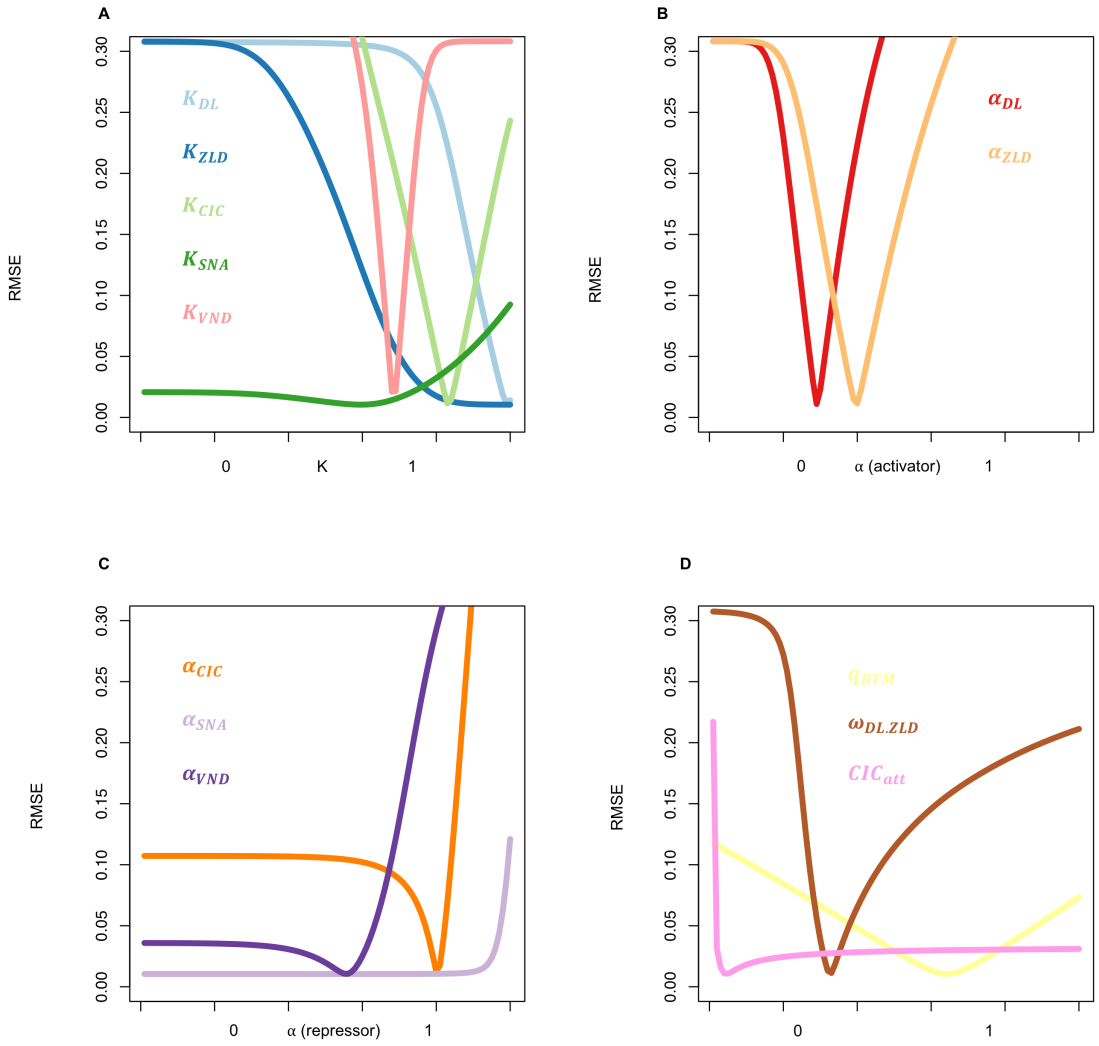

Figure A.4: Sensitivity plots for a model (MST) corresponding to A.3A Figure. (A-D) Panels show the RMSE scores of the model as the corresponding parameter's value is varied within its range, keeping other parameters fixed at their optimized values.

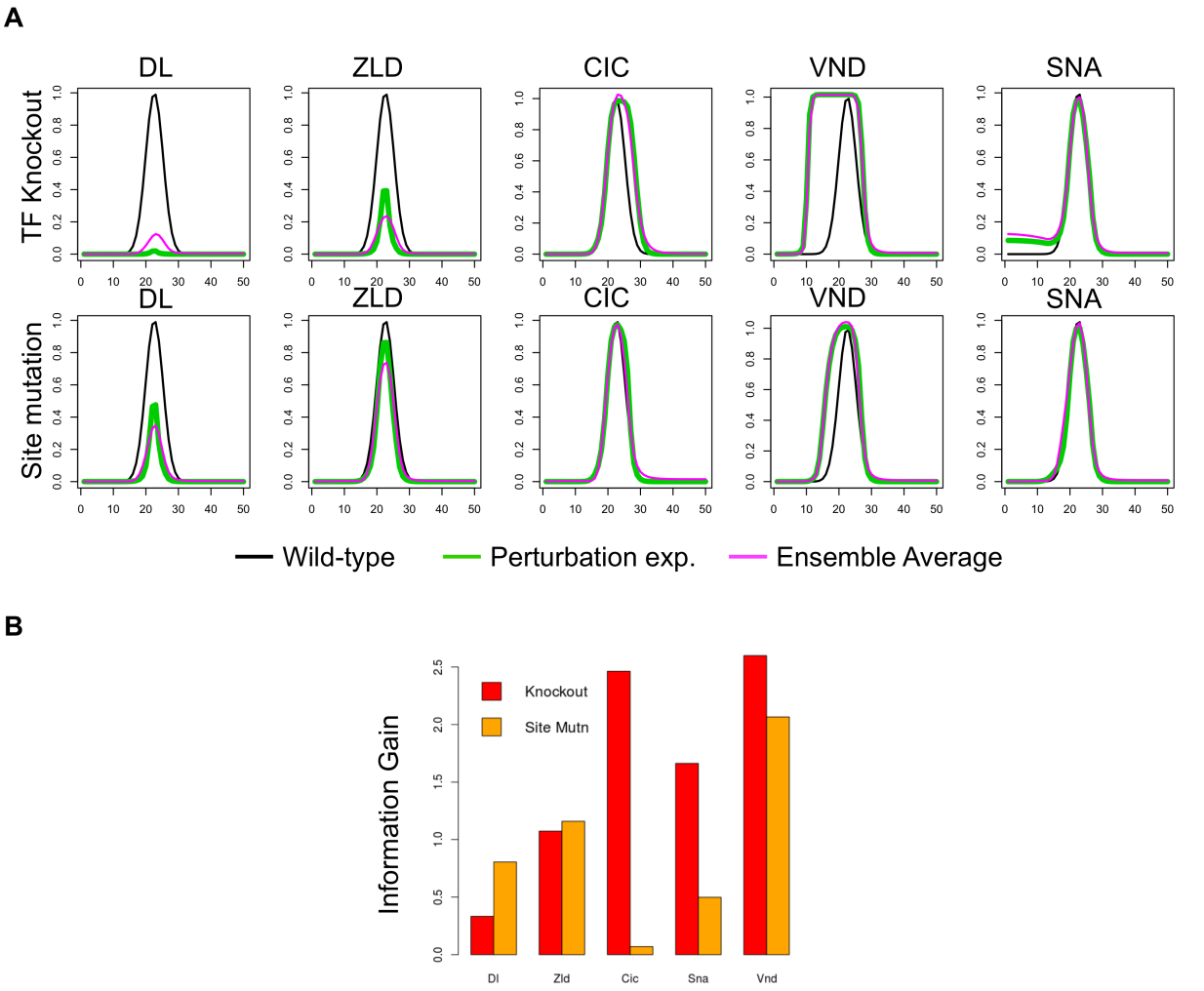

Figure A.5: Evaluating in silico experiments with an alternative 'synthetic real' model MST (A.3B Fig) that is distinct from that used in Figure 2.2. (A) The model is used to generate synthetic 'experimental' results of TF knockout (top row) or strongest site mutagenesis (bottom row), for each TF, shown in green. These are compared to the synthetic 'wild-type' expression profile of ind, shown in black (in each panel). Magenta curves show the average prediction of the filtered ensemble for each of these 'experiments'. (B) Information gain due to each synthetic perturbation experiment, with semantics analogous to those in Figure 2.2 B, under the alternative 'synthetic real' model MST.

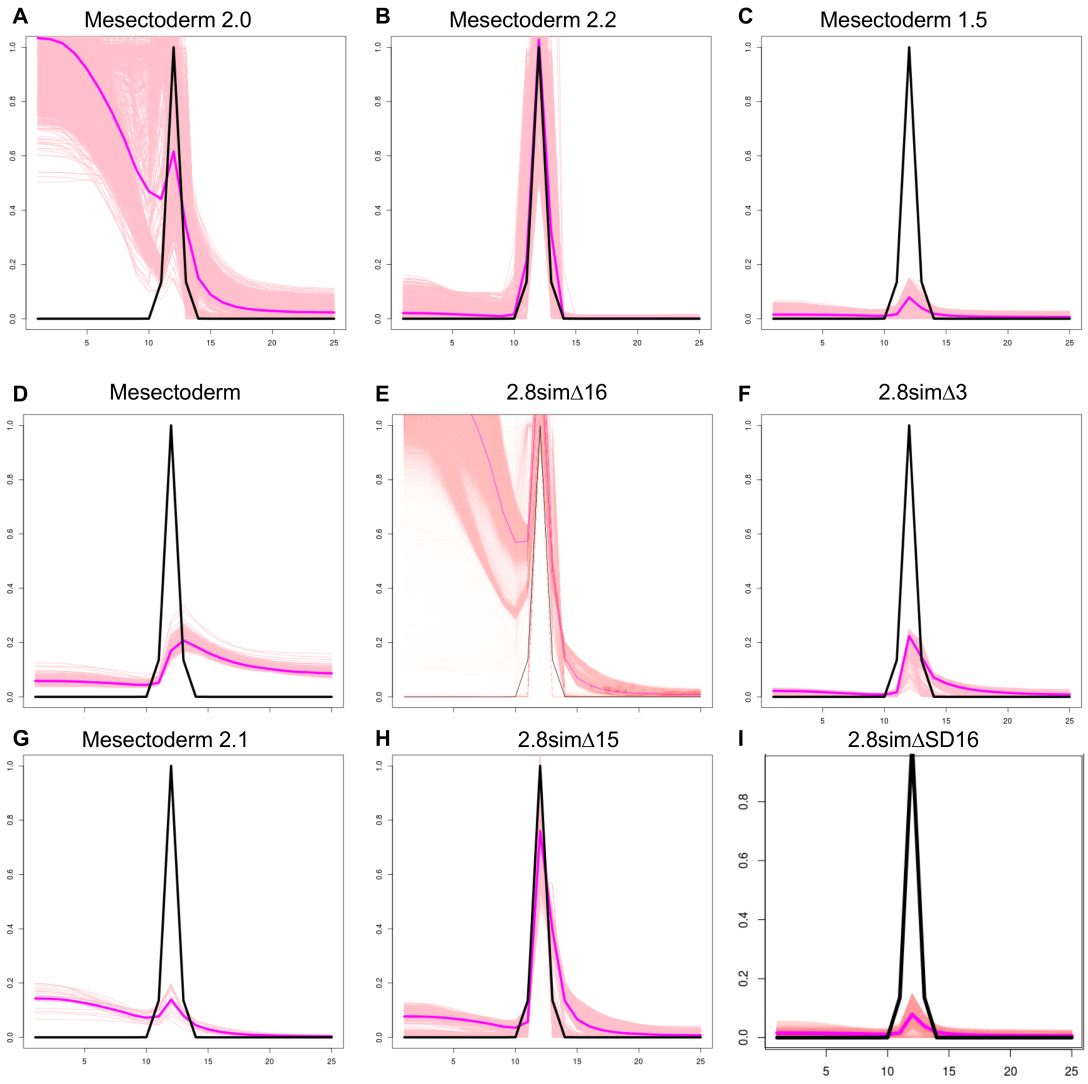

Figure A.6: Filtered ensembles for sim perturbation experiments. The wild-type ensemble was filtered to retain only models that correctly predict the observed effects of an additional perturbation experiment (A.2 Table), thus yielding a smaller, filtered ensemble for each experiment. In each panel, wild-type sim expression is shown in black, pink curves represent predictions of models in the ensemble and magenta represents a weighted average of the ensemble predictions. (A) The expression extends to the presumptive mesoderm (B) No change is observed in the expression when the strongest DL site is mutated. (B) Wild-type expression was observed in mesectoderm. (C) The expression is abolished when using the 1.5 Kb enhancer sequence. (D) The ventral-most line of cells of the neurogenic ectoderm. Weak and variable staining is also detected in more ventral regions of early embryos. (E) Weak expression ("greatly reduced mesectodermal transcription, but a low level of expression was detectable"). (F) Mesectodermal transcription was abolished. (G) No expression is observed. (H) The expression is similar to the wild-type expression. (I) The Expression is completely abolished.

Table A.1: Perturbation experiments for ind gene, from the literature. For each perturbation experiment reported in the literature, second column summarizes the effect on ind expression and the third column reports a criterion that we selected, based on the observed effect, for determining if a model's prediction is consistent with that experiment. We used these criteria to filter the wild-type ensemble of models. Note that expression profiles are described with the D/V axis being divided into 50 bins, with the ventral-most position being bin 1 and the dorsal-most position being bin 50.

Experiment (Source Pubmed ID)	Observation	Filtering Criteria for Model Prediction
DL 1 site mut. (Strongest site mutagenized.) (22216201)	No change is observed.	Predicted expression pattern has at most 5% error with SSE measure compared to the same models prediction on the wild-type sequence.
DL 3 sites mut. (Three overlapping sites removed.) (27136354)	Peak expression is reduced by 65%.	Expression is low in all bins outside bins number 22-28 (average expression less than 0.01 of the peak). Peak expression is less than 40% of the wild-type level.
ZLD site mut. (Four strongest sites removed.) (27136354)	Expression reduced to half of the endogenous levels.	Expression is low in all bins outside bins number 22-28 (average expression less than 0.01 of the peak). Peak expression is less than 60% of the wild-type.
CIC site mut. (Site mutagenized.) (23733957)	Expression domain expands dorsally, where it matches the spatial domain of the DL protein.	Average expression in bins 40-50 is less than 5% and that in bins 25-35 is greater than 80% of the maximum expression.
SNA KO (SNA knockout.) (16750631)	No change is observed.	Predicted expression has less than 5% SSE from the same model's prediction in wild-type conditions.
VND KO (VND knockout.) (9832511)	Expression expand ventrally, beyond the peak of VND mesoderm region.	Average expression in bins 1-10 is less than 5% of the wild-type and it is more than 80% of the wildtype in bins 20-25.

Table A.2: Perturbation experiments for sim gene, from the literature. For each perturbation experiment reported in the literature, second column summarizes the effect on sim expression and the third column reports a criterion that we selected, based on the observed effect, for determining if a model's prediction is consistent with that experiment. We used these criteria to filter the wild-type ensemble of models. Note that expression profiles are described with the D/V axis being divided into 50 bins, with the ventral-most position being bin 1 and the dorsal-most position being bin 50. The SSE score is evaluated on the first 25 bins.

Experiment (Source Pubmed ID)	Observation	Filtering Criteria for Model Prediction
2.8sim (9840810)	Wild-type expression is observed in mesectoderm in one row of cells on either side of the embryo.	The peak of expression is at bin 14 and is more than 0.8 in scale of 1. The average expression in bins 13-15 is more than 0.3 and the average expression is less than 1% in all the other bins.
mesectoderm2.1 (9840810)	No expression	Predicted expression profile differs from a flat line of no expression by an SSE score of less than 5%.
mesectoderm1.5 (9840810)	No expression	Predicted expression profile differs from a flat line of no expression by an SSE score of less than 5%.
2.8$sim\Delta$3 (9840810)	No expression ("mesectodermal transcription was abolished.")	Peak of predicted expression profile is less than 25% of the wild-type peak.
2.8$sim\Delta$16 (9840810)	Weak expression ("greatly reduced mesectodermal transcription, but a low level of expression was detectable")	The average expression in bins 1-9 is more than 50%of the peak in predicted expression profile.
2.8$sim\Delta SD$16 (9840810)	No expression ("completely abolished mesectodermal transcription")	Predicted expression profile differs from a flat line of no expression by an SSE score of less than5%.
2.8$sim\Delta$15 (9840810)	Mesectoderm ("did not affect mesectodermal transcription")	Predicted expression profile differs from wild type profile by SSE score of less than 5%.
mesectoderm (15128669)	The ventral-most line of cells of the neurogenic ectoderm. Weak and variable staining is also detected in more ventral regions of early embryos.	Average expression in bins 10-15 is greater than 10% of the peak expression and the average expression in bins 1-9 is less than 5%.

Table A.3: Information gain due to a perturbation experiment following another experiment. The value in row i and column j is the information gain due to experiment j when conducted after experiment i.

	DL 1 site mut.	DL 3 sites mut.	ZLD site mut.	CIC site mut.	SNA KO	VND KO
DL 1 site mut.	0	0.90	1.60	1.63	2.66	2.01
DL 3 sites mut.	0.86	0	1.43	1.06	1.85	1.69
ZLD site mut.	0.69	0.55	0	1.18	1.43	1.26
CIC site mut.	0.46	-0.07	0.93	0	0.49	0.22
SNA KO	0.76	0	0.46	-0.23	0	-0.19
VND KO	0.34	0.06	0.51	-0.27	0.04	0

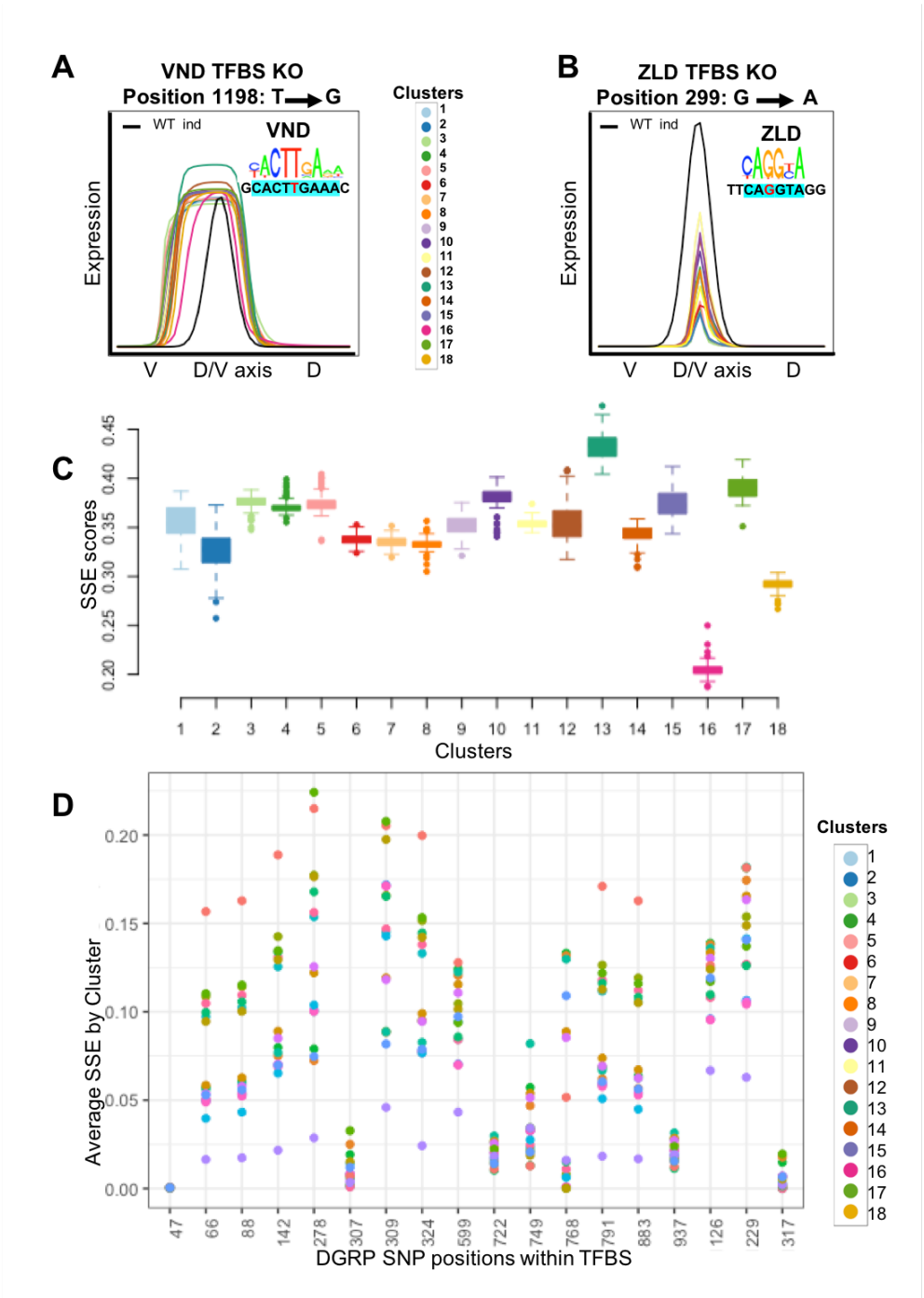

Figure A.7: (A) Gene expression profile of ind enhancer predicted by different models that cluster together. Each cluster or group of models is depicted by a separate color and the plotted expression profile is the average expression predicted from models in that cluster. (B) The mutation selected in this part hits a perfect ZLD site and has the second largest predicted impact. (C) Distribution of the sum of squared error (SSE) between wild-type expression and predicted expression for the mutation shown in (A), across models in each cluster. (D) Average SSE of predictions made by each cluster of models, for each SNP in DGRP population.

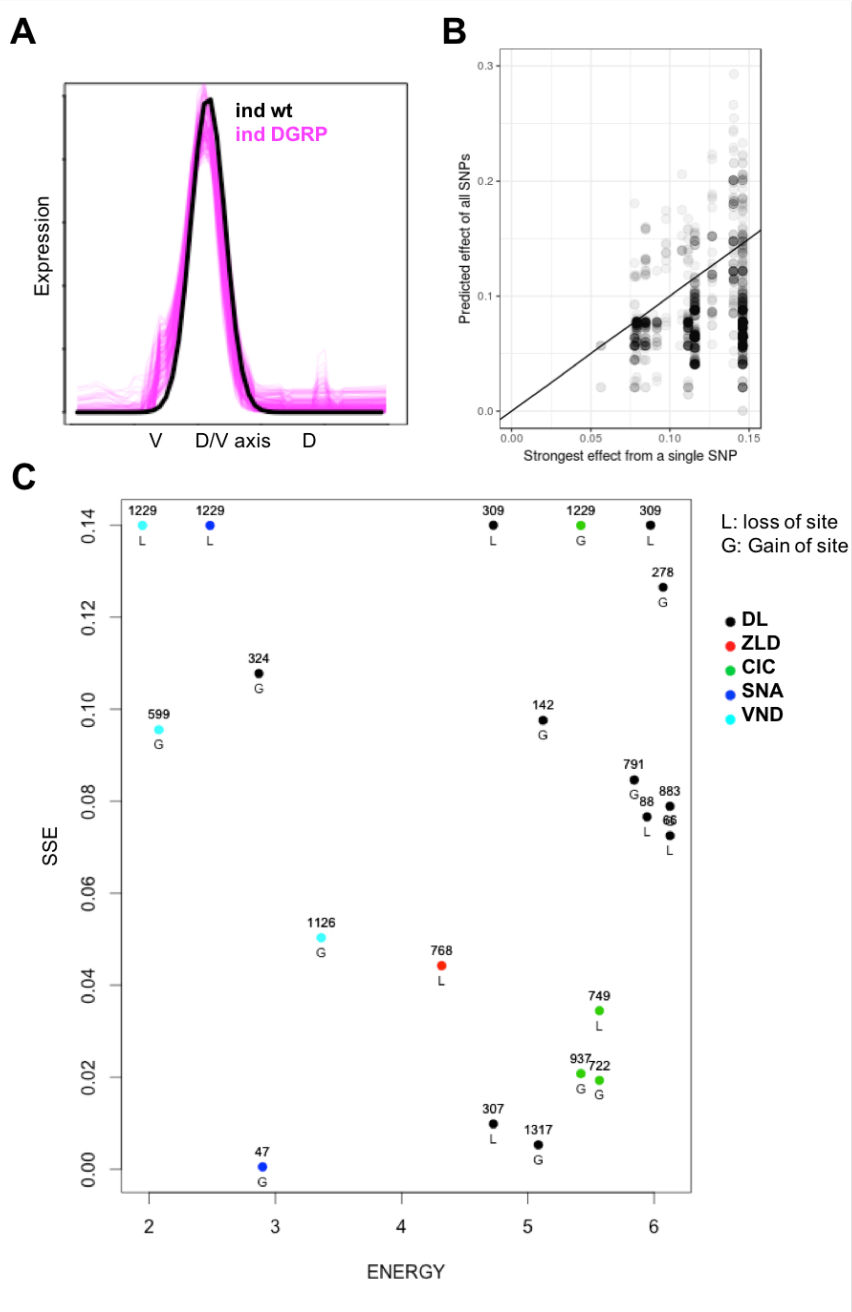

Figure A.8: A) Predicted expression profile of individuals in DGRP population is close to the wild-type expression of ind gene. Predictions are averages over the entire ensemble of models. (B) Simulated individuals with the same population-wide allele frequencies as DGRP population do not exhibit compensatory effect of mutations as observed in DGRP population. Each individual is a point on the scatter plot. The x-axis is the SSE score between the predicted gene expression profile of the individual and the wild-type expression. The y-axis is the SSE for a construct that includes only the strongest-effect SNP present in the individual. (C) Each point in the scatter plot is a SNP from DGRP. The x-axis shows the predicted strength of the TF binding site that overlaps with the SNP and the y-axis is the change in SSE score between the model-predicted wild-type and the mutation.

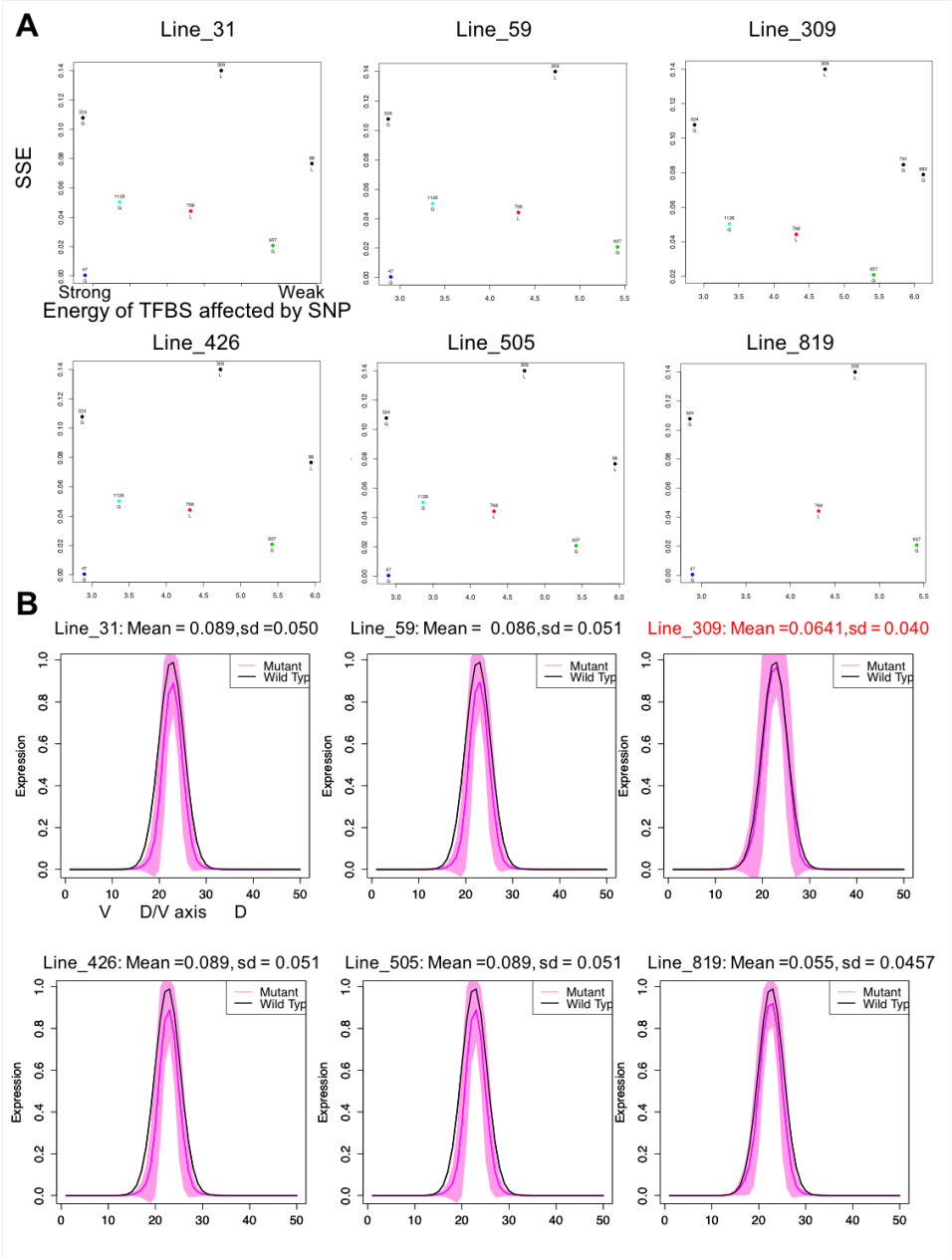

Figure A.9: (A) Each panel corresponds to an individual in DGRP population. Similar to figure S2(D), points are the SNPs that the individual carries, colored by the TF whose predicted binding site overlaps the SNP. The x-axis shows the predicted strength of this binding site and the y-axis is the SSE score between the model predicted wild-type and the mutation. (B) Gene expression profile of the ind enhancer genotype for DGRP individuals, as predicted by the models (pink), compared to known expression profile of wild-type enhancer (black). Each panel's title shows the mean and standard deviation (sd), across models, of SSE scores between the model-predicted profile for the enhancer sequence defined by the genotype of a particular individual and the wild-type expression.

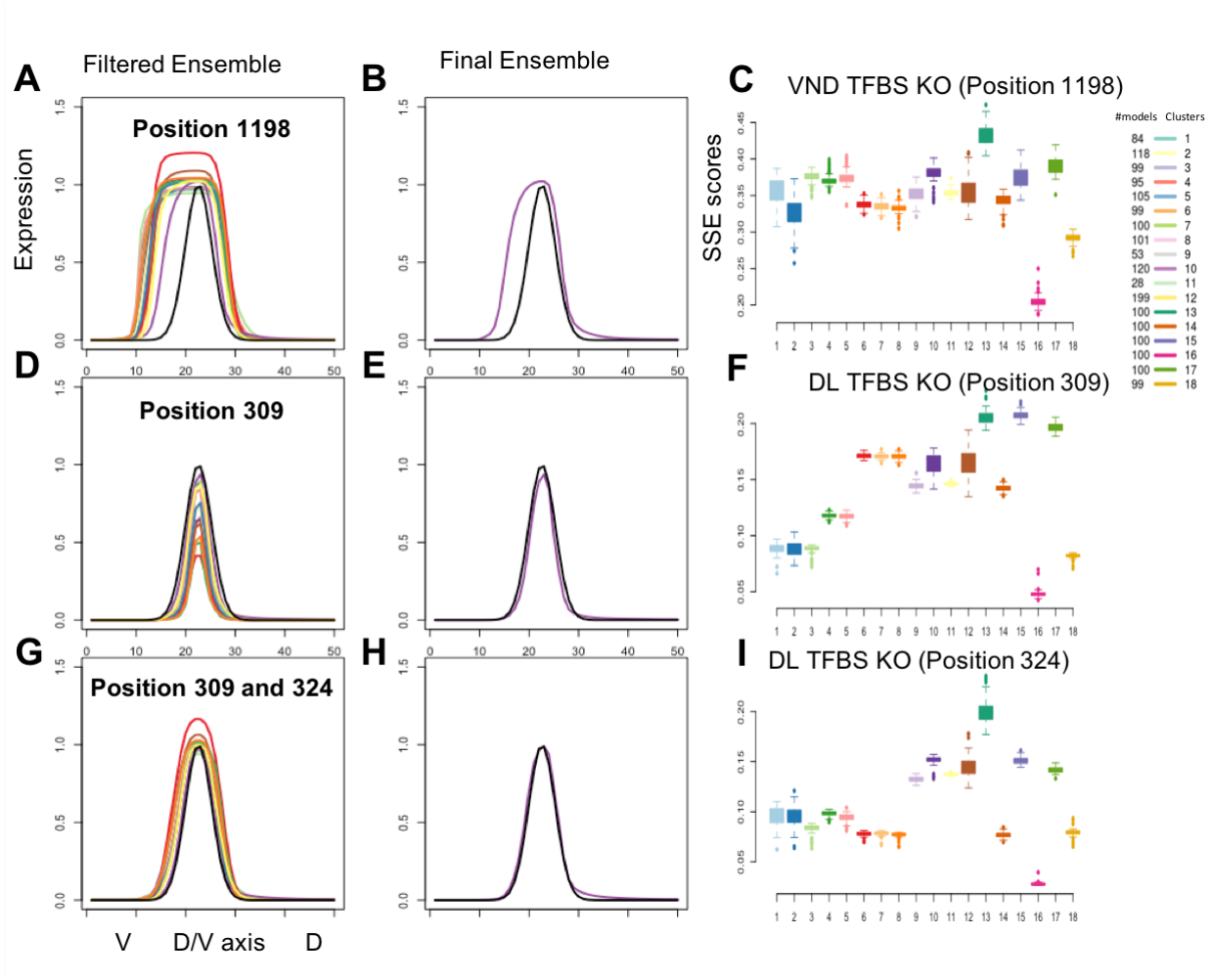

Figure A.10: (A) Model-predicted expression profile of "construct1", the first construct selected for the experiments, where the SNP selected for the construct lies within a VND binding site. Each color represents the average prediction of a group of models that cluster together in the parameter space. (B) The expression profile predicted by a select group of models ('final ensemble') suggests little impact of the SNP on expression. (C) SSE score (between wild-type expression and model-predicted expression for the construct) for each group of models and the number of models in that group. The magenta boxplot, with the lowest SSE scores, corresponds to the final ensemble. (D) Similar to A, but for "construct2"; this construct carries a mutation in DL binding sites and most models predict an impact on the gene expression. (E) The final ensemble predicts no change in the gene expression for this construct. (F) The SSE score of the magenta cluster is the lowest. (G) Similar to A and D, but for "construct3"; this construct carries two mutations in DL binding sites that are potentially compensatory. Most models predict no change in the gene expression for this construct. (H) The final ensemble predicts no change in the gene expression for this construct. (I) Similar to C and F, the magenta cluster has the lowest SSE scores among all groups of models.

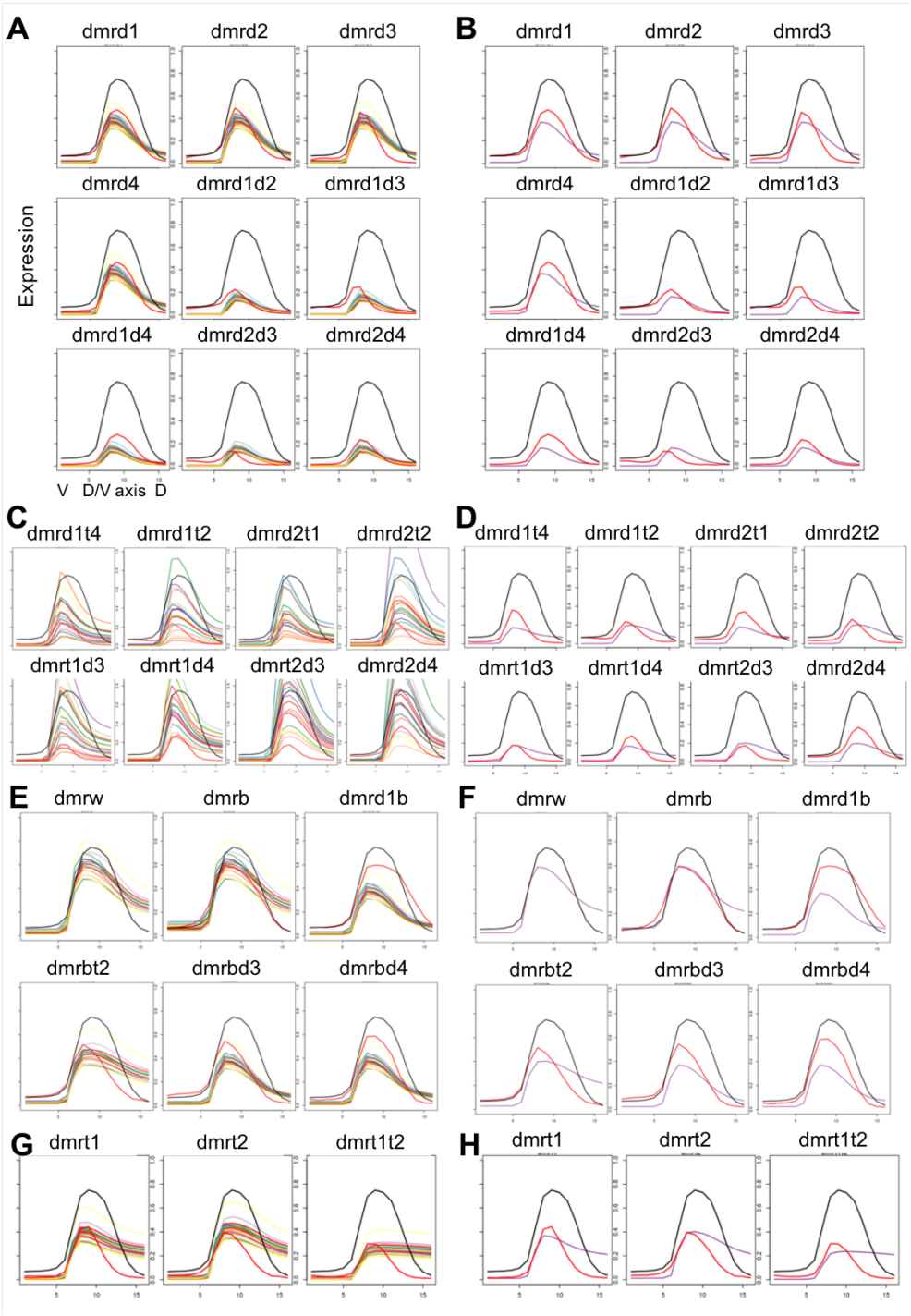

Figure A.11: Each panel shows wild-type rho profile (black), expression profile of the synthetic enhancer (name provided in panel title) (red), and expression profile for that enhancer as predicted by one or more clusters of models. (A, C, E, G) Predicted expression profiles correspond to each cluster of models. (B, D, F, H) Predicted expression profile represents the special cluster of models called the final ensemble. Enhancers in (A, B) represent DL site mutagenesis, (C, D) represent DL and TWI site mutagenesis, (E, F) represent BHLH site mutagenesis and (G, H) represent TWI site mutagenesis.

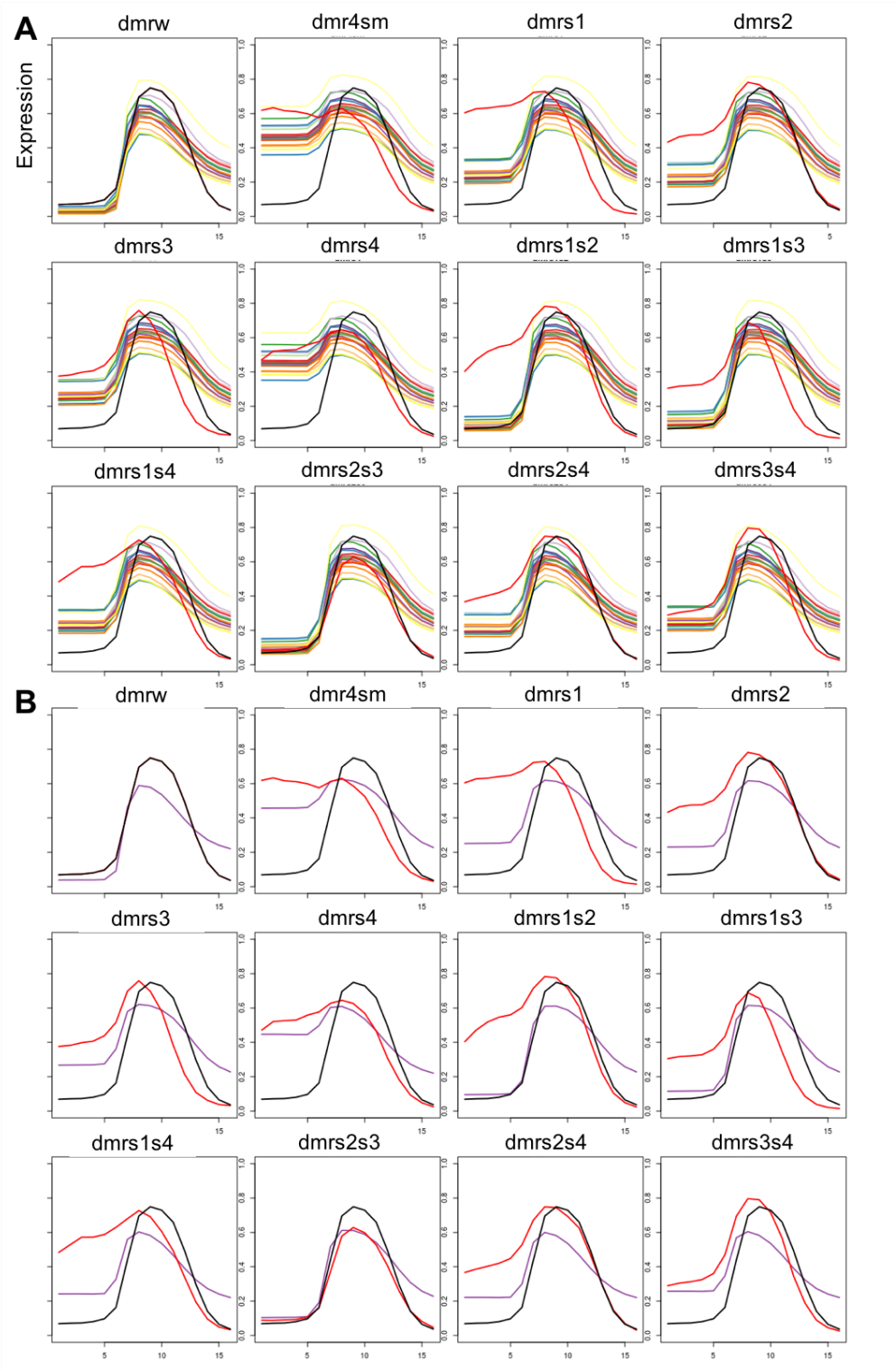

Figure A.12: Similar to Figure S5, except that these represent synthetic variants of the rho enhancer where SNA binding sites have been mutagenized. Wild-type expression profile (black) and experimental expression profile of the synthetic enhancer (red) are compared to model-predicted profiles from (A) each cluster of models or (B) only the models in the final ensemble.

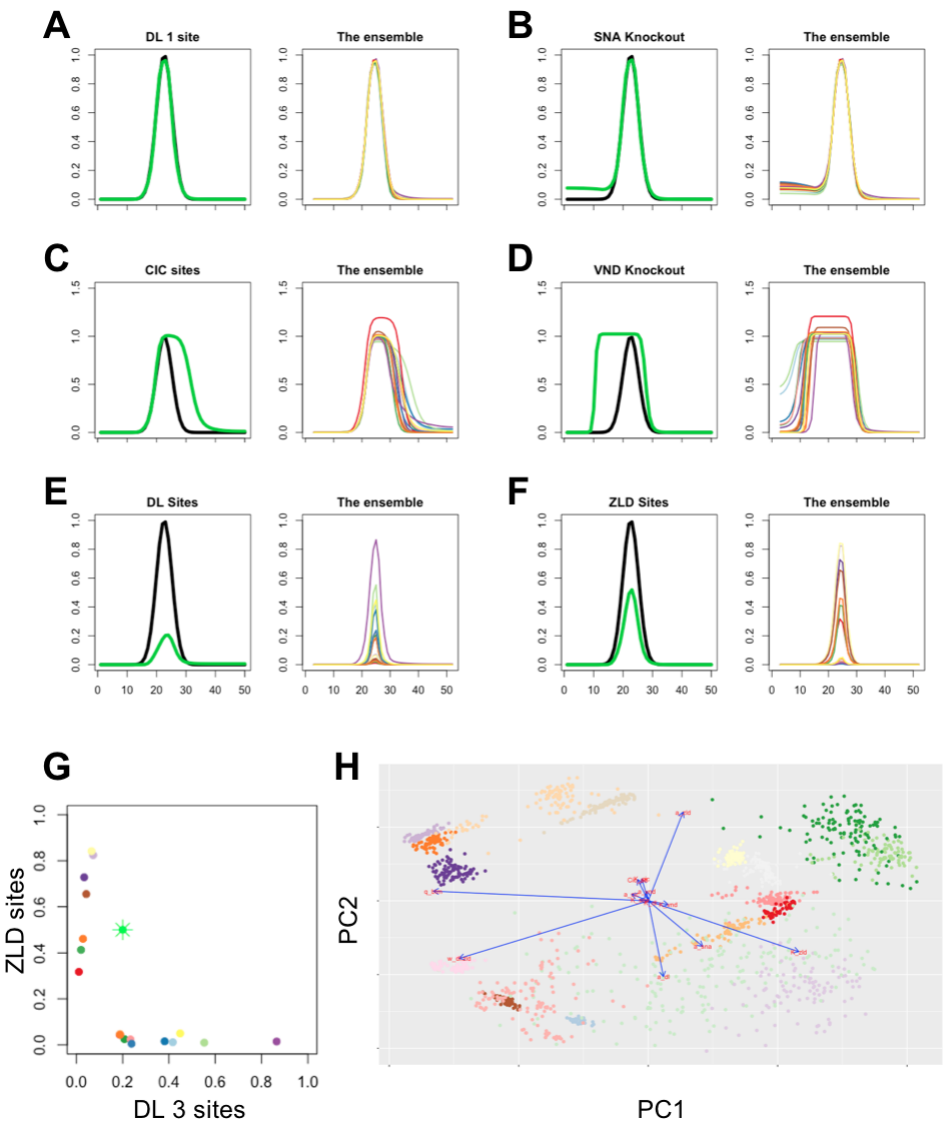

Figure A.13: (A-F) Models trained on wild-type ind enhancer are filtered based on their ability to predict the effects of perturbations. Expression profiles in wild-type and perturbation conditions are shown in black and green respectively, and predictions made by each cluster of models in the ensemble are shown in other colors in adjacent panels. (G) Twelve models survive 5 of the 6 filters depicted in A-F, but no model simultaneously predicts the effect of perturbations denoted by "ZLD sites" (mutagenesis of four ZLD sites) and "DL 3 sites" (mutagenesis of three DL sites). The peak expression levels in these two conditions are shown by the green star, while other points represent peak expression levels in these two conditions as predicted by the 12 above-mentioned models. The x-axis is the peak expression level (real or predicted by a model) for the "DL sites" mutagenesis condition and the y-axis is the peak expression level for the "ZLD sites" mutagenesis condition. (H) Regions of parameter space around the 12 above-mentioned models are sampled densely. Shown is a PCA-based two-dimensional representation of these models.

Table A.4: Liver tissue validations. The number of overlaps from the enrichment tests. The size of all eQTL set in Liver is 207822. The number of overlaps with H3K27ac broad peaks is 9877 and narrow peaks is 5408. The number of eRNAs overlap is 453. The first four rows of the table reports the actual number of overlaps and the last three rows are the $-\log_{10} p$ where p is the p-value of the hyper-geometric test.

Liver	baseline	PGM	Expected	baseline	PGM	Expected
#Genes	76	517	94	614	1249	239
#SNPs	100	100	100	995	1000	1000
#TFs		15	15		19	19
#H3K27ac-broad	24	30	4.75	209	230	47.52
#H3K27ac-narrow	14	8	3.58	95	59	35.76
H3K27ac-broad-Pvals	11.22	16.62	0.42	73.37	89.39	0.34
H3K27ac-narrow-Pvals	7.28	2.91	0.77	29.71	8.37	1.60
eRNA-Pvals	0.71	5.51	0.22	2.84	20.78	3.06

Table A.5: Colon tissue validations. The size of all eQTL set is 417338. The number of overlaps with H3K27ac broad peaks is 30462 and narrow peaks is 13577. The number of eRNAs overlap is 779.

Colon	baseline	PGM	Expected	baseline	PGM	Expected
#Genes	46	175		774	1155	
#SNPs	100	100	100	1000	1000	1000
#TFs		14	20		20	20
H3K27ac-broad	23	25	7.23	239	256	72.99
H3K27ac-narrow	16	7	3.25	143	96	32.53
eRNA	0	0	0.19	2	17	1.87
H3K27ac-broad-Pvals	6.77	8.04	0.39	60.12	70.78	0.32
H3K27ac-narrow-Pvals	7.59	1.78	0.16	40.08	20.26	0.34
eRNA-Pvals	0.94	0.94	0.76	0.54	11.82	0.49

Table A.6: Heart tissue validations. The size of all eQTL set is 471757. The number of overlaps with H3K27ac broad peaks is 21896 and narrow peaks is 14884. The number of eRNAs overlap is 848.

Heart	baseline	PGM	Expected	baseline	PGM	Expected
#Genes	65	207		718	1200	
#SNPs	100	100	100	995	1000	1000
#TFs		14	20		20	20
H3K27ac-broad	25	26	5.25	217	266	52.47
H3K27ac-narrow	21	16	3.57	179	93	35.64
eRNA	0	0	0.20	1	2	2.03
H3K27ac-broad-Pvals	12.29	13.17	0.55	80.82	121.12	0.79
H3K27ac-narrow-Pvals	12.20	7.78	0.57	78.62	19.60	3.75
eRNA-Pvals	0.78	0.78	0.98	0.78	0.57	0.58

Table A.7: Lung tissue validations. The size of all eQTL set is 775469. The number of overlaps with H3K27ac broad peaks is 31233 and narrow peaks is 20016. The number of eRNAs overlap is 1185.

Lung	baseline	PGM	Expected	baseline	PGM	Expected
#Genes	46	170		774	1571	
#SNPs	100	100	100	1000	1000	1000
#TFs		11	19		19	19
H3K27ac-broad	25	26	4.02	348	343	40.28
H3K27ac-narrow	13	4	2.58	172	93	25.81
eRNA	0	0	0.15	1	3	1.53
H3K27ac-broad-Pvals	13.70	14.64	0.43	219.78	214.25	0.34
H3K27ac-narrow-Pvals	6.50	0.93	0.46	85.77	25.50	0.35
eRNA-Pvals	0.85	0.85	1.00	0.35	1.16	0.52